One-shot learning and eligibility traces in sequential decision making

THÈSE N^{O} 9191 (2018)

PRÉSENTÉE LE 18 DÉCEMBRE 2018
À LA FACULTÉ DES SCIENCES DE LA VIE
LABORATOIRE DE CALCUL NEUROMIMÉTIQUE (IC/SV)
PROGRAMME DOCTORAL EN NEUROSCIENCES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Marco Philipp LEHMANN

acceptée sur proposition du jury:

Prof. J. Gräff, président du jury
Prof. W. Gerstner, Prof. K. Preuschoff, directeurs de thèse
Prof. C. Summerfield, rapporteur
Dr J. Gläscher, rapporteur
Dr M.-O. Gewaltig, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

Wer A sagt, der muß nicht B sagen.
Er kann auch erkennen, daß A falsch war.
— Bertolt Brecht

Für Olivia, Elin & Jérôme

Acknowledgements

Doing a Ph.D. in neuroscience, nine years after finishing a master in computer science, is exciting, challenging, unusual, fulfilling, rewarding . . .
I would like to thank my two advisors Prof. Kerstin Preuschoff and Prof. Wulfram Gerstner for giving me this opportunity, and for their support and inspiration throughout the past years.

He Xu and Marc Lauffs from the Lab of Psychophysics helped me with the pupillometry experiments, while Marc Repnow solved our technical problems. Inspiration for the neuroeconomic aspects of my project came from Kerstin Preuschoff's team in the "Neuroscience in Finance and Economics" Lab. In particular, I would like to thank Leyla Loued-Khenissi for helpful comments.

The work on sequential decision making was an inspiring collaboration with Vasiliki Liakoni. Dane Corneil and Johanni Brea provided critical comments and useful advice.
Working in the LCN is a privilege. Discussions with lab members were always inspiring and helpful. For me, the most unique trait of the LCN team is the amazing flexibility in switching from in-depth discussions (not only during lab meetings) to comical and sometimes even absurd debates on virtually any topic (not only at Sat). The people that made up this ever changing, unique cocktail of ideas and opinions are (around 2017/18): Aditya Gilra, Alex Seeholzer, Bernd Illing, Chiara Gastaldi, Dane Corneil, Florian Colombo, Hesam Setareh, Ho Ling Li, Johanni Brea, Laureline Logiaco, Mohammadjavad Faraji, Olivia Gozel, Samuel Muscinelli, Tilo Schwalger and Vasiliki Liakoni.

Die wichtigste Unterstützung erhielt ich von Christa. *Mit unendlicher Energie hast Du tausend Dinge in Balance gehalten, die vielen Veränderungen mit getragen sowie unsere Kinder Jérôme, Elin und Olivia unterstützt. Gemeinsam erlebten wir eine facettenreiche Zeit in Lausanne. Danke!*

Und ich bedanke mich bei meinen Eltern. *Ihr gabt uns auf dem Erlenhof die Wurzeln, das Vertrauen und die Freiräume um in viele Richtungen wachsen zu können.*

Lausanne, 28 September 2018 M.L.

Abstract

When humans or animals perform an action that led to a desired outcome, they show a tendency to repeat it. The mechanisms underlying learning from past experience and adapting future behavior are still not fully understood. In this thesis, I study how humans learn from sparse and delayed reward during multi-step tasks.

Learning a sequence of multiple decisions, from a reward obtained only at the end of the sequence, requires a mechanism to link earlier actions to later reward. The theory of reinforcement learning suggests an algorithmic solution to this problem, namely, to keep a decaying memory of the state-action history. Such memories are called eligibility traces. They bridge the temporal delay between the moment an action is taken and a subsequent reward. We ask whether humans make use of eligibility traces when learning a sequential decision making task. The difficulty in answering this question is that different competing algorithmic solutions make similar predictions about behavior. Only during a few initial trials, learning with eligibility traces is qualitatively different from other algorithms.

Here, I implemented a novel learning task with an experimental manipulation that allowed us to guide participants through a controlled sequence of states. With this hidden manipulation, we were able to isolate the specific trials in which the competing models are distinguishable. Behavioral data as well as simultaneously recorded pupil dilation revealed effects compatible with eligibility traces, but not with simpler models. Furthermore, the trial-by-trial reward prediction errors were correlated with pupil dilation and EEG measurements.

Our experimental data show effects of eligibility traces in behavior and pupil data, after a single experience of state-action associations, which has not been studied before in a multi-step task. We view our results in the light of one-shot learning and as a signature of a learning mechanism present both in temporal difference and one-shot learning.

Keywords human learning, sequential decision making, eligibility trace, one-shot learning, credit assignment

Zusammenfassung

Menschen und Tiere neigen dazu belohntes Verhalten zu wiederholen. Die Lernmechanismen, welche dazu führen, Verhalten aufgrund vergangener Erfahrungen anzupassen, sind noch nicht vollständig verstanden. In dieser Dissertation untersuche wie Menschen lernen, wenn sie in mehrstufigen Entscheidungsexperimenten gelegentlich belohnt werden.

Um optimales Verhalten zu erlernen, wenn eine Belohnung erst am Ende einer Sequenz von Entscheidungen erfolgt, verlangt nach einem Mechanismus, welcher die früheren Entscheidungen mit der später erfolgten Belohnung in Verbindung bringt. Die Theorie des verstärkenden Lernens (engl. Reinforcement Learning) schlägt zur Lösung dieses Problems einen algorithmischen Ansatz vor, nämlich von den Zustands-Entscheidungs-Paaren eine abklingende Erinnerung (engl. eligibility trace) im Speicher zu halten, um die Zeitspanne von der Entscheidung bis zum Erhalt einer Rückmeldung zu überbrücken. In dieser Arbeit stellen wir die Frage, ob Menschen solche Erinnerungen verwenden, wenn Sie ein sequentielles Entscheidungsaufgabe erlernen. Die Schwierigkeit in der Beantwortung dieser Frage besteht darin, dass alternative Lösungsansätze ähnliche Voraussagen über das Verhalten machen. Lediglich während einer frühen Phase des Lernens sind die verschiedenen Lösungen unterscheidbar. Wir haben ein neues Lernexperiment entwickelt, welches uns erlaubt die Studienteilnehmer durch bestimmte Sequenzen zu führen und so, die entscheidenden Bedingungen herbeizuführen, unter welchen wir unsere Frage beantworten können. Sowohl die Verhaltensdaten als auch die gleichzeitig erhobenen Pupillenmessungen zeigten Signale, welche mit der Hypothese übereinstimmten, nicht jedoch mit alternativen Modellen. Ausserdem konnten wir die theoretischen Lernsignale sowohl mit den Pupillen- als auch mit EEG-Daten korrelieren. In unseren experimentellen Daten zeigen sich die Effekte einer Lernstrategie mit abklingenden Erinnerungen bereits nach einer einzigen Kopplung. Dies konnte bisher in mehrstufigen Entscheidungsexperimenten nicht gezeigt werden. Wir interpretieren unsere Ergebnisse im Kontext von schnellem Lernen (engl. one-shot learning) und sehen diese als Hinweise dafür, dass abklingende Erinnerungen sowohl beim verstärkenden Lernen als auch beim schnellen Lernen eine Rolle spielen.

Contents

Acknowledgements **v**

Abstract (English) **vii**

List of figures **xii**

List of tables **xiii**

1 Introduction **1**

1.1 Learning and decision making 2
  1.1.1 Sequential decision making 2
  1.1.2 The credit assignment problem 2
  1.1.3 Credit assignment heuristics in human learning 3
1.2 Reinforcement learning theory 4
  1.2.1 Markov Decision Process 4
  1.2.2 Solving an MDP: finding the optimal policy 5
  1.2.3 Temporal Difference Learning 7
  1.2.4 Eligibility traces and credit assignment 9
  1.2.5 Neo-Hebbian learning in a simple neuronal model 9
1.3 The neuroscience of human learning and decision making 11
  1.3.1 A short history of the reward prediction error hypothesis of dopamine . 11
  1.3.2 A model of value-based learning with eligibility traces 15
  1.3.3 Multiple learning systems and the tole of one-shot learning 18
1.4 Pupil dilation: a proxy for neural activity 20
1.5 Thesis contribution 22

2 Evidence for behavioral eligibility traces in human learning **23**

2.1 Abstract 23
2.2 Introduction 25
2.3 Results 27
  2.3.1 Task design and experimental conditions 27
  2.3.2 Evidence for eligibility traces: Delayed reward reinforces state-action associations 29
  2.3.3 Pupil dilation changes at states which received delayed reward 29

2.3.4 Models with eligibility trace explain behavior better than alternative models 33
2.3.5 Modeling quantifies an eligibility trace time-scale of 10 seconds 34
2.3.6 Reward prediction errors at non-goal states modulate pupil and EEG . . 37
2.4 Discussion 39
2.5 Supplementary Materials and Methods 41
2.5.1 Experimental conditions 41
2.5.2 Pupil data processing 42
2.5.3 Action assignment in the Markov Decision Process 42
2.5.4 Reinforcement Learning Models 44
2.5.5 Parameter Fit and Model Selection 45
2.5.6 Regression Analysis 47
2.5.7 EEG Recordings and Analysis 48
2.5.8 Parameter fit and model selection: supplementary results 50

3 Discussion 53
3.1 One-shot learning and eligibility traces 53
3.1.1 One-shot learning and episodic control in human experiments 53
3.1.2 Evidence for one-shot learning and the role of eligibility traces 54

A Human learning and sequential decision making. An fMRI Study 57
A.1 Abstract 57
A.1.1 Introduction 59
A.2 Methods 60
A.2.1 Experimental design to separate MF from MB prediction errors 60
A.2.2 Reinforcement Learning algorithms 61
A.2.3 Data Analysis 62
A.3 Results 63
A.3.1 De-correlation of RPE and SPE 63
A.3.2 Behavior: Actions close to goal are learned faster 63
A.3.3 Learning algorithms and behavioral fit 66
A.3.4 Neural signatures of multiple learning algorithms 66
A.4 Discussion 69

B Adaptive control between multiple reinforcement learning systems 71

Bibliography 102

Curriculum Vitae 103

List of Figures

1.1 Reinforcement Learning: Interaction between agent and environment 5
1.2 Temporal Difference Learning and predictions for experiments 6
1.3 Hebbian and neo-Hebbian learning 10
1.4 Cortico - Basal Ganglia - Thalamo - Cortical Loop 16
1.5 Dopamine and Norepinephrine Projections 17
1.6 Multiple systems for learning and control 19

2.1 Experimental design 28
2.2 Delayed reward reinforces state-action associations 30
2.3 Immediate and delayed reward affect pupil dilation in subsequent state visits . 31
2.4 Eligibility for reinforcement decays with a time-scale τ in the order of 10 seconds 35
2.5 Reward prediction error (RPE) at non-goal states modulates pupil dilation . . . 38
2.6 Pupil dilation during control experiment 43
2.7 Parameter fitting Q-λ 46
2.8 Regression analysis reveals temporal profile of RPE modulation 47
2.9 Detailed results of regression analysis and permutation tests 49
2.10 EEG Data Analysis: RPE - ERP correlation 50
2.11 Sarsa-λ parameter fit per participant 51
2.12 Correlation of SARSA Parameters 52

A.1 Sequential decision making task with surprise trials 61
A.2 Learning curves per state 64
A.3 Post-hoc validation of the RPE/SPE de-correlation 65
A.4 Models and model selection 67
A.5 Neural correlates of RPE and SPE 68

1 Introduction

Learning from past experience and adapting future behavior is an impressive capacity of the brain. For example, when starting to learn how to ride a bicycle, most trials fail. But with repetition, the behavior adapts and improves until the muscles perform the right movements automatically. This kind of learning is colloquially called *trial and error* learning.

There is a long tradition in behavioral psychology and neuroscience to study how humans and animals acquire behavior after observing positive or negative outcomes of their actions. In control theory and engineering, similar topics have been studied: how can an agent (e.g. a robot), guided by a reward signal, learn to take correct actions in specific situations? The two domains formalized the study of learning into *Reinforcement Learning*.

In this thesis, I present my research in human learning and sequential decision making. We designed a maze-like task to study how humans learn, through trial and error, the sequence of states and actions that led them to a goal state. The question we ask is how humans integrate this sparse and delayed feedback to guide future behavior. We discuss our experimental results with respect to two aspects. First, we explain behavior in terms of classic reinforcement learning algorithms, with a focus on the role of eligibility traces. Then, we discuss our experimental results in the light of one-shot learning.

This introduction is divided into four sections. In the first section, I will introduce the problem of sequential decision making. Section 1.2 is a short review of the computational framework of reinforcement learning. In section 1.3 I turn to neuroscience; in a short, chronological review, I highlight the milestones that have led to the *reward prediction error hypothesis of dopamine*. This is followed by the review of a model of learning, which is thought to implement aspects of reinforcement learning in the brain. In the last part of this section I introduce the role of one-shot learning for behavior. Finally, section 1.4 explains the link between neural activity and pupil dilation.

1.1 Learning and decision making

1.1.1 Sequential decision making

Sometimes life is easy and a simple action leads to immediate pleasure; eating chocolate is such situation. But most of the time, a rewarding outcome is the consequence of a, potentially long, sequence of actions. Walking to the next library to borrow an interesting book, baking homemade bread, or winning a Go game are such examples. In fact, at the level of motor control, even eating chocolate is a complex sequence of fine-tuned muscle activations. The study of sequential decision making tries to understand how such behavior is learned.

Psychologists have identified a rich set of factors that influence behavior and learning, such as motivational and emotional states. Here we adopt a simplified view, and reduce the driving forces to the abstract goal of *maximizing reward*. That is, we consider the case of an agent seeking to perform the sequence of actions that will maximize the reward eventually obtained.

1.1.2 The credit assignment problem

When feedback is received only at the end of a long sequence of actions, how does a learner know which actions were responsible for producing the outcome? This is known as the *credit assignment problem* [Minsky, 1961].

A classic example to illustrate the *credit assignment problem* is chess. In each turn, a player has many possible options to choose from. Some moves may increase the chances to win, while others reduce it. The difficulty of learning the good moves (or avoiding the bad ones) is that none of the intermediate decisions receive direct feedback; there is only the final outcome, win, lose or stalemate.

The chess example and a possible solution were discussed in the 1950s by Allen Newell in "The chess machine: an example of dealing with a complex task by adaptation" [Newell, 1955]:

> The problem of sample-size requires mention: How large a sample of experience is necessary to obtain learning? Or better: How much information about the effects of behavior is necessary to successfully modify the behavior? Chess affords a good example of this problem. It is extremely doubtful whether there is enough information in "win, lose, or draw" when referred to the whole play of the game to permit any learning at all over available time scales. There is too much behavior. For learning to take place, each play of the game must yield much more information. This is exactly what is achieved by breaking the problem into components. The unit of success is the goal. If a goal is achieved, its subgoals are reinforced; if not, they are inhibited. (Actually, what is reinforced is the transformation rule that provided the subgoal.) This is so whether the game is ultimately won or lost.

Newell's suggestion of *breaking the problem into components* is an example of hierarchical (or *divide and conquer*) approaches to complex problems. While this is a proven engineering approach, a biological system would have to solve a new problem: that of identifying (or learning) the sub-problems themselves. This is not impossible, and there is evidence for hierarchical reinforcement learning [Botvinick et al., 2009, Ribas-Fernandes et al., 2011] and hierarchical planning [Balaguer et al., 2016] in humans.

But there are also simpler, heuristic approaches. Imagine a sequential decision making task where the last action made leads to the rewarding goal state. It is reasonable to assume that not only the last action caused the rewarding outcome, but also the next to last (and second to last, etc.) action has contributed to success. This intuition is captured by the temporal proximity heuristic: Credit should be assigned not only to the action which was immediately followed by reward, but also to earlier actions.

1.1.3 Credit assignment heuristics in human learning

If a choice is followed by a reward, then humans and many animals tend to repeat the same choice in the future. While such repetition biases have been observed in many experiments, there are more subtle effects:

In a classic experiment, Thorndike [Thorndike, 1933] observed a *spread of effect*. That is, he observed a repetition bias not only for choices that were immediately followed by a reward, but also for the choices made in close temporal proximity, prior to a reward.

In a recent study, Jocham et al. [2016] also observed this spread of effect and identified mechanisms driving it. In their experiment, human participants had to learn stimulus-reward associations for three different visual stimuli (geometric shapes) from two forms of feedback: a delayed, causal reward, and an instantaneous, random reward. In addition to the expected learning of causal stimulus-reward associations, behavior also revealed *non-causal* learning, driven by two different heuristic mechanisms: First, choices were repeated if they were in temporal proximity, prior to a reward. Second, "subjects were likely to assign credit for a reward to a choice that was frequently selected in the recent past, whether or not it was causally related to the reward" [Jocham et al., 2016].

The two effects can be seen as credit assignment heuristics based on recency and frequency. These approaches have also been proposed and formalized in machine learning research as solutions to the *credit assignment problem*, as described in the next section.

1.2 Reinforcement learning theory

Reinforcement Learning (RL) theory studies the puzzling problem of a hedonic agent that *wants* something, and *learns* how to get it through interaction with its environment.

Starting without knowledge about the environment, an agent explores the environment by taking actions. These actions can initially be random, or follow a more elaborated exploration strategy. At some point, the choices lead to a reward and the agent uses this information to receive more reward in the future: it adapts its future behavior. Many trials might be necessary to explore a dynamic environment and discover optimal actions: reinforcement learning algorithms describe iterative processes of adapting the behavioral strategy. Figure 1.1 shows an early version of a still influential abstraction of the Reinforcement Learning problem [Minsky, 1961].

Reinforcement learning theory covers two aspects: first, it formally describes the problem statement, and second, it provides solutions to solve it.

We start with the description of the problem statement.

1.2.1 Markov Decision Process

A sequential decision making problem with observable states and stochastic transitions can be described as a Markov Decision Process (MDP). An MDP is a mathematical object composed of the following five elements:

S: the set of states.
A: the set of actions.
$P^a_{ss'}$: The probability to transition from state s to s' when taking action a.
$R^a_{ss'}$: The immediate reward after the transition from state s to s' when taking action a.
γ: Distal reward discount factor in $[0,1]$ [1]

The state transitions $P^a_{ss'}$ satisfy the Markov property: the distribution over the next states s' depends only on the current state s, not on the history of previous states. We only consider the case of finite sets of discrete states and actions and discrete-time dynamics.

The behavior of an agent in such an environment is defined by its *policy* π. $\pi(s,a)$ is the probability to take action a in state s.

Starting from some state s and following a policy π, the interaction between an agent and the environment (the MDP) produces a sequence of states and actions. Such a sequence can be infinite (inifinite horizon MDP) or episodic, meaning that there are absorbing states which terminate an episode.

---

[1] When studying learning, we consider γ as a property of the learning agent, instead of the MDP.

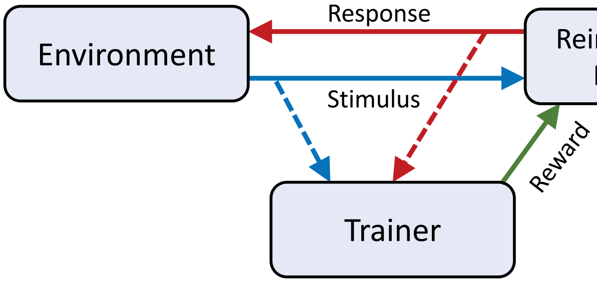


Fig. 1.1. Figure adapted from the classic "Steps Toward Artificial Intelligence" Minsky [1961]. This early version of a still valid abstraction of the RL problem shows the interactions between the environment and an agent (reinforcement machine). Interestingly, here the reward is not "returned by the environment" as it is often described in more recent texts, but generated by a separate part of the system, akin to a *critic*. Original caption: "Note that the Trainer need not know how to solve problems, but only how to detect success or failure, or relative improvement".

1.2.2 Solving an MDP: finding the optimal policy

When an agent selects actions according to an *optimal* policy π^*, it will maximize the expected reward. The goal of solving an MDP is to find π^*. There is a wide range of algorithms for solving an MDP [Sutton and Barto, 2018]. Most algorithms are based on estimating an intermediate quantity, the so-called value function, which depends on the policy π. The value $V^{\pi}(s)$ of a state s, is an estimate of how much reward to expect when starting in s and select actions according to the behavioral policy π. This is illustrated in Figure 1.2[a2], which depicts the state values of a simple MDP example. Similarly, action values $Q^{\pi}(s,a)$ quantify the expected reward after selecting action a in state s and following π for the later steps.

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \tag{1.1}$$

Where t is the current time step, and r_{t+k+1} is the reward obtained when transitioning from s (in a future time step $t+k$) to s' (in $t+k+1$). The parameter $\gamma \in [0,1]$ is the discount factor that controls the weight of distal rewards.

These Q-values guide behavior. A common choice for selecting actions given the values of the available options, is the *softmax policy*:

$$\pi(s,a) = \frac{exp(Q(s,a)/T)}{\sum_{\tilde{a}} exp(Q(s,\tilde{a})/T)} \tag{1.2}$$

The temperature T controls the balance between *exploration* (for large T, actions are chosen with equal probability) and *exploitation* (for $T \rightarrow 0$, the action with the largest Q-value is chosen).

The definition (Eq. 1.1) can be transformed into a recursive relationship, known as the *Bellman*

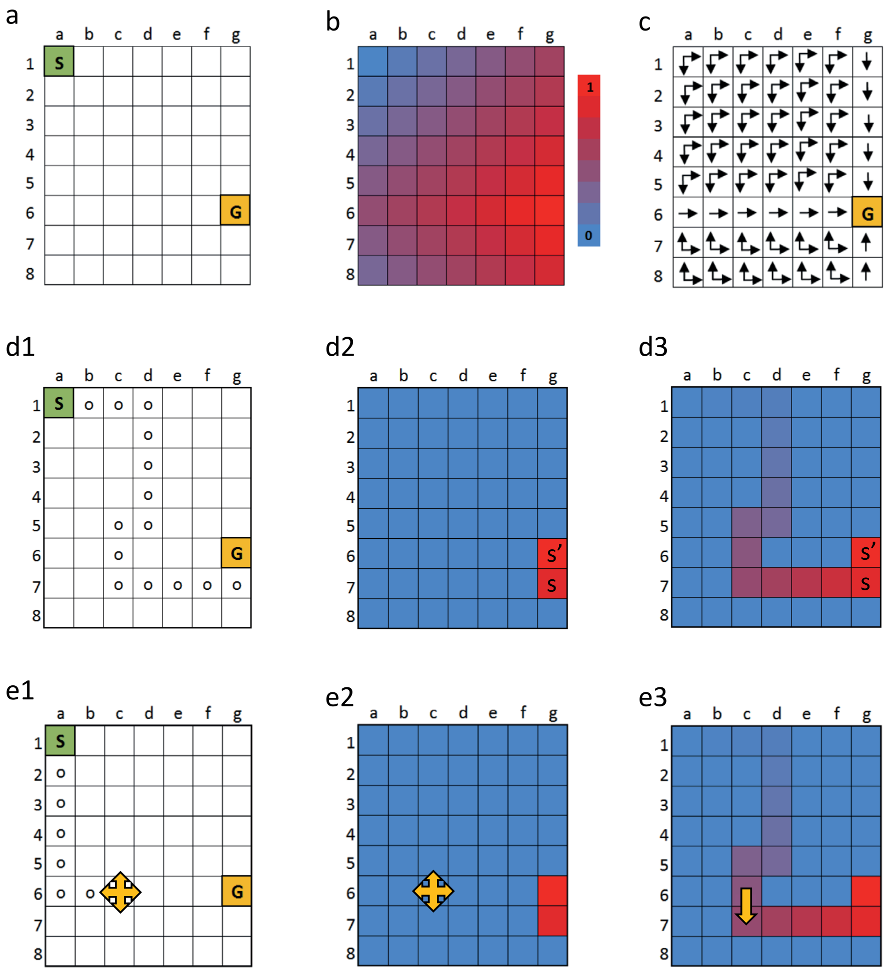


Fig. 1.2. Temporal Difference Learning: example and predictions for experiments
[a] Gridworld is a simple MDP example. It is defined by 64 states, 4 actions (up, down, left, right) in each state, deterministic transitions, a single reward state G, $\gamma < 1$, and a single start state S. **[b]** Value function V(s). States with high value are colored in red, low values in blue. If an agent has access to this function, the task is solved; the policy can be derived from it. **[c]** Policy π^*. Arrows indicate the optimal action in each state.
Initially, an agent does not have access to the value function. It has to learn it through interaction with the environment. **[d1]** Episode one. An agent discovers the goal state after random exploration. **[d2]** The value function was initialized with V(s)=0 for all states (blue). Discovering the goal state generates a large prediction error. TD-0 learning updates V(g7) only. **[d3]** The decaying eligibility traces allow TD-λ to update the values of visited states.
[e1] Episode two. The agent reaches state c6. **[e2]** A TD-0 learner has no information to guide the decision at c6. **[e3] A TD-λ update in episode one makes two predictions about episode two: i) the transition from b6 to c6 causes a reward prediction error, and ii) the agent has a bias to select the action "down". In chapter 2 we test these two predictions in a simpler task with humans.**

expectation equation. It decomposes the value into an immediate reward and a (discounted) expected future reward.

$$\begin{aligned} Q^{\pi}(s,a) &= E_{\pi}\{\, r_{t+1} + \gamma Q^{\pi}(s',a') \quad | s_{t+1} = s', a_{t+1} = a', s_t = s, a_t = a\} \\ &= \sum_{s'} P^{a}_{s,s'} \left[R^{a}_{s,s'} + \gamma \sum_{a'} \pi(s',a') Q^{\pi}(s',a') \right] \end{aligned} \tag{1.3}$$

For a given policy π and a known MDP, that is if one has access to $P^{a}_{s,s'}$ and $R^{a}_{s,s'}$, one can solve the system of linear equations (Eq. 1.3) to find the state-action values Q^{π}. Here we consider a different case: an agent that wants to maximize rewards is interested in finding an optimal policy π^* for which the Q-values are maximal:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{1.4}$$

The optimal state-action values satisfy the nonlinear *Bellman optimality equation*:

$$Q^*(s,a) = \sum_{s'} P^{a}_{s,s'} \left[R^{a}_{s,s'} + \gamma \max_{a'} Q^*(s',a') \right] \tag{1.5}$$

To solve the nonlinear optimality equation 1.5, iterative methods known as *Value Iteration* and *Policy Iteration* can be used. A discussion of these dynamic programming algorithms can be found in Sutton and Barto [2018]. These algorithms need a fully specified MDP and fall into the category of *model-based*[2] RL.

In many interesting and realistic problems an agent does not know the dynamics of the environment. It could still try to estimate $P^{a}_{s,s'}$ and $R^{a}_{s,s'}$ through interaction with the environment and apply a model-based algorithm to infer action values. Alternatively, there are *model-free* algorithms which learn how to act without knowing or estimating $P^{a}_{s,s'}$ and $R^{a}_{s,s'}$. The next section discusses such a learning algorithm.

1.2.3 Temporal Difference Learning

From the many algorithms, solving Eq. 1.5, we now consider one particular class: temporal difference (TD) learning.

Model-free methods improve the estimates of Q-values iteratively from experience without estimating the dynamics of the environment, but by averaging over observed rewards: At any moment t during learning, the agent has some estimate $Q_t(s,a)$ of the expected future reward. This estimate could be improved by rolling out a full episode, observe all the collected rewards,

---

[2] The term *model* refers to the dynamics of the environment, specified by $P^{a}_{s,s'}$ and $R^{a}_{s,s'}$.

and update the estimate of $Q_t(s, a)$ accordingly.

Instead of doing a full roll-out, a TD agent considers only a *single* action a_t, the observed next state s_{t+1}, and reward r_{t+1} for the Q-value update [3]. TD learning is based on two properties of the Bellman equation: First, it decomposes the expected reward $Q_t(s_t, a_t)$ into immediate reward r_{t+1} and expected future reward $Q_t(s_{t+1}, a_{t+1})$. Based on this property a TD learner does a single step, observes only r_{t+1} and uses $Q_t(s_{t+1}, a_{t+1})$ to estimate the remaining future rewards. Second, the Bellman equation is a consistency equation that holds *after* convergence. During learning, the equality does not hold. Instead, for a single sample, and using estimates of Q-values, we have the following equation:

$$Q_t(s_t, a_t) + \delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) \tag{1.6}$$

The right hand side of Eq. 1.6 forms a target value: after learning, $Q_t(s_t, a_t)$ correctly predicts this target, and δ goes (on average) to 0.

The quantity δ is defined as the *reward prediction error*. It absorbs two sources of errors: first, Q_t are only estimates for the true Q-values. The goal of learning is to remove this part of the error. Second, Q-values express expectations, while here, by sampling, we observe concrete realizations of the potentially stochastic rewards and transitions. This part of the error reflects the intrinsic uncertainty of the environment and can not be removed. Estimating this variance to optimize learning in uncertain environments, is an interesting research field, both, computationally and experimentally [Nassar et al., 2012, Payzan-LeNestour et al., 2009, Behrens et al., 2007], but not further discussed in this thesis.

A classic TD learning algorithm, SARSA, uses the trial-by-trial reward prediction error δ to update the Q-value estimates:

$$\begin{aligned} \delta_t &= r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \\ Q_{t+1}(s, t) &\leftarrow Q_t(s, t) + \alpha \delta_t \end{aligned} \tag{1.7}$$

Where $\alpha \in [0, 1]$ is the learning rate.

So far we have discussed how Q-values are updated. But how does an agent find the optimal policy? The answer is surprisingly simple (but proving it is difficult [Singh et al., 2000]): the Q-values and the policy depend on each other (see Eq. 1.2 for the case of the softmax family) and SARSA converges *simultaneously* to Q^* and π^* under appropriate conditions (see [Singh et al., 2000, Sutton and Barto, 2018]).

---

[3] Using the estimate $Q_t(s_{t+1}, a_{t+1})$ instead of the actual future rewards is called *bootstrapping*. It introduces a bias but reduces variance.

1.2.4 Eligibility traces and credit assignment

The update rule in Eq. 1.7 changes only the value of one state-action value, $Q_t(s_t, a_t)$. That is, credit for the prediction error δ is assigned only to the action immediately preceding it. As discussed in 1.1.2, actions further in the past have also contributed to the outcome and could be updated.

This is implemented in SARSA-λ, a modification of SARSA that keeps track of the past state-action activations in a table $e_t(s,a)$. Each time action a_t at state s_t is taken, the eligibility for this state-action pair to be updated is set to 1. For all other state-action pairs, this eligibility decays by a factor λ:

$$e_t(s,a) = \begin{cases} 1 & \text{if } s = s_t, a = a_t \\ \gamma\lambda e_{t-1}(s,a) & \text{otherwise.} \end{cases} \tag{1.8}$$

This form of *eligibility traces* is known as replacing traces and implements the recency based credit assignment heuristic [Singh and Sutton, 1996]. The update rule (Eq. 1.7) becomes

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha\delta_t e_t(s,a) \tag{1.9}$$

The update with and without eligibility traces is illustrated in Figure 1.2 for the case of V-values. Figure 1.2[c] also illustrates how learning with eligibility traces makes predictions about the reward prediction error and about the behavior. We will come back to these two predictions in more detail in chapter 2, where we experimentally test them using a tailored task.

1.2.5 Neo-Hebbian learning in a simple neuronal model

The SARSA algorithm can be easily implemented in a computer, using any programming language. Before discussing, in section 1.3, how model-free reinforcement learning is thought to be implemented in the complex circuits of the human brain, it is useful to think about a very basic neuronal circuit.

Figure 1.3 depicts two models: they both implement the simple functionality of activating an action neuron a when a state neuron s is active. They differ in how this state-action associations are strengthened: In classic Hebbian learning (Fig. 1.3[a]) the co-activity of the pre- and the postsynaptic neurons leads to a (small) weight change. This unsupervised method exploits the statistical correlations, but ignores the reward information. The strength of the synapses develops slowly over many trials.

In three factor rules of synaptic plasticity, the co-activity of pre- and postsynaptic neurons does not lead to an immediate weight change, but marks the synapse as eligible for change. Only in the presence of a third factor, this eligibility leads to an actual weight change (Fig. 1.3[b]). The reward signal differentiates co-activity in unsuccessful trials from rewarded trials. This allows for larger learning rates than in the unsupervised case, and results in faster learning.

Three factor learning rules, and experiments providing support for the role of eligibility traces in synaptic plasticity, are reviewed in detail in Gerstner et al. [2018].

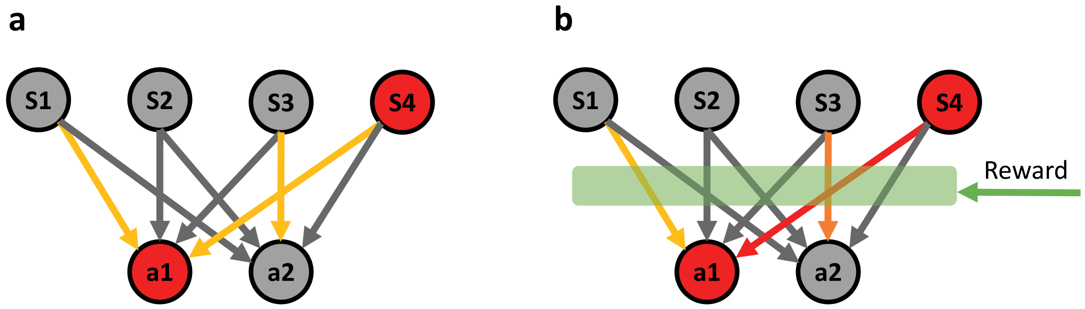


Fig. 1.3. Hebbian and neo-Hebbian learning Neurons S1 to S4 represent states, neurons a1 and a2 two possible actions. Active neurons are marked in red. Arrows indicate an excitatory projection from the state to the action neurons. The synaptic weight change after a sequential co-activation of (S1, a1) - (S3, a2) - (S4, a1), followed by reward is shown. Colors symbolize the weight change (from strong to weak: red, orange, yellow; no change is marked as gray). **[a]** In classical models of Hebbian learning, the co-activity of the pre- and postsynaptic neuron leads to an immediate weight change. Accordingly, the co-activity of (S1, a1) has strengthened the synaptic efficacy from S1 to a1. Similarly (S3, a2) and (S4, a1) associations are strengthened. The weight change is relatively weak. In correlation-based learning, strong associations are formed over many repetitions. For learning rewarded behavior, neo-Hebbian learning is more efficient: **[b]** Illustration of neo-Hebbian (or three factor) learning with eligibility traces. Each state-action co-activation has left a decaying memory at the corresponding synapse. This eligibility for weight-change is turned into a strengthening (or weakening) of the synapse only at the arrival of the third factor (e.g. a dopamine mediated reward prediction error). Note the specificity of learning: although the reward is signaled globally, only synapses with a trace of past co-activation are actually changed. The reward guided gating allows for larger weight changes.

1.3 The neuroscience of human learning and decision making

Summing up numbers, denoted as r_t, is all an abstract agent, described in the previous section, is aiming at. This does not seem to be a very rich model for the fascinating and sometimes (seemingly) irrational behaviors of humans. Not surprisingly, the neural circuits behind human behavior are complex and far from being fully understood. What *is* surprising, is that the simple temporal difference learning models can, to some extent, explain human learning and that there is a close correspondence between the reward prediction error and the activity of dopaminergic midbrain neurons, known as the *Reward Prediction Error Hypothesis of Dopamine*. In this section I will briefly summarize the experimental and conceptual milestones that have led to this discovery. I will then discuss how the human brain is thought to implement reinforcement learning.

1.3.1 A short history of the reward prediction error hypothesis of dopamine

This compact review is mostly based on material from Colombo [2014], Gazzaniga et al. [2008] and Sutton and Barto [2018].

Edward Thorndike's *law of effect* is considered one of the first general statements about the strengthening (or weakening) of an association between stimuli and behavior [Gazzaniga et al., 2008, Beeler, 2012].

> If an association is followed by a satisfying state of affairs it will be strengthened and if it is followed by an annoying state of affairs it will be weakened. [Thorndike, 1911]

Here, the *strengthening* of an association refers to an observable repetition bias in future behavior. That is, behavior becomes, to a certain extent, predictable.

An important conceptual step forward in the study of learning was made when Bush and Mosteller [1951] formalized learning as a form of *error correction*, an idea further developed by Rescorla and Wagner: "Organisms only learn when events violate their expectations. Certain expectations are built up about the events following a stimulus complex; expectations initiated by the complex and its component stimuli are then only modified when consequent events disagree with the composite expectation." [Rescorla et al., 1972]. The Rescorla-Wagner learning rule describes how an initially neutral stimulus becomes predictive (conditioned) of a subsequent reward (or unconditioned stimulus). The rule quantifies the strengthening of such an association as a trial-by-trial change, proportional to how strongly a stimulus predicted subsequent reward.

While this rule explained many experimental results, it failed explaining second-order conditioning [Glimcher and Fehr, 2013]: In second-order conditioning, an animal first learns to associate a stimulus S1 with a reward. After this first phase of learning, a second stimulus S2 is associated with S1. Second-order conditioning is the phenomenon where the animal learns to associate S2 with reward indirectly through the pairing with S1 only.

The Rescorla-Wagner rule fails explaining the conditioning of S2 because during the S2-S1 pairing, no reward (or unconditioned stimulus) is involved, and therefore the association strength between S2-S1 does not change. In other words, the rule only considers one-step experiments and explains learning from *immediate*, but not from *delayed* reward. A solution to this problem, namely temporal-difference (TD) learning, was introduced by R. Sutton in 1988 [4]. The proposed algorithm learns from reward prediction errors at rewarded and non-rewarded states: “The hallmark of temporal-difference methods is their sensitivity to changes in successive predictions rather than to overall error between predictions and the final outcome.” [Sutton et al., 1988].

In retrospect, it may seem obvious how the Rescorla-Wagner rule generalizes into temporal-difference learning. But one should note an important conceptual difference between the two: The Rescorla-Wagner rule quantifies the *associative strength* between a previously neutral stimulus and an unconditioned stimulus. On the other hand, TD learning tackles the entangled aspects of multi-step learning, which include value prediction and credit-assignment. When introducing TD learning, R. Sutton writes in "Learning to Predict by the Methods of Temporal Differences" 1988:

> This article introduces a class of incremental learning procedures specialized for prediction that is, for using past experience with an incompletely known system to predict its future behavior. Whereas conventional prediction-learning methods assign credit by means of the difference between predicted and actual outcomes, the new methods assign credit by means of the difference between temporally successive predictions. Although such temporal-difference methods have been used in Samuel’s checker player, Holland’s bucket brigade, and the author’s Adaptive Heuristic Critic, they have remained poorly understood. [Sutton et al., 1988].

As discussed in section 1.2.3, TD learning builds on two basic concepts, restated here for the case of V-values: First, the *reward prediction* $V(s)$ associated with each state s, quantifies the expected future reward when starting in s. Second, the *reward prediction error* RPE is the difference between reward predictions of two successive states s_t and s_{t+1}:

$$RPE(t) = r_{t+1} + V(s_{t+1}) - V(s_t) \tag{1.10}$$

The Reward Prediction Error (RPE) is the crucial learning signal. It drives learning through the

---

[4] Sutton built his theory on earlier ideas, notably [Samuel, 1959] and Klopf [1972]. Later, he discovered an earlier version of TD-learning: Witten [1977]. See Sutton and Barto [2018] for more details.

following update of the reward prediction:

$$V(s_t) \leftarrow V(s_t) + \alpha RPE \tag{1.11}$$

With learning, the V-values approach reliable predictions, while the RPEs decrease.

The link between the theoretical model of TD-learning and neurophysiological data was established in the years between 1991 and 1997. According to Colombo [2014], in 1991 P. Dayan and R. Montague recognized a signature of TD-learning in recordings from dopamine neurons in monkeys, published by W. Schultz, but the publication “was getting rejected by every major journal in neuroscience”. It was not until 1996 that they were able to publish their work [Montague et al., 1996], and in 1997, the seminal article "A neural substrate of prediction and reward" [Schultz et al., 1997] was published. There, the authors observed (and modeled) four characteristic firing patterns in dopaminergic neurons: 1) high activity at the delivery of an unpredicted reward, 2) baseline activity at the delivery of a predicted reward, 3) high activity at the onset of a reward predicting stimulus, 4) below-baseline activity when a predicted reward was *not* delivered. This confluence of computational theory with behavioral and neurophysiological data, is now known as the *Reward Prediction Error Hypothesis of Dopamine*.

Soon after, and enabled by the emerging fMRI technology, neural correlates of reward prediction errors were identified in the human brain [Berns et al., 2001, Pagnoni et al., 2002, O’Doherty et al., 2003, McClure et al., 2004, Pessiglione et al., 2006]. Also, using EEG, Holroyd and Coles [2002] identified characteristic neural signatures of human error processing and proposed a model for adaptive behavior. Since that time, a wealth of research has advanced the understanding of the neural circuits underlying learning and value-based decision making (for reviews see Doya [2008], Niv [2009], Glimcher [2011], Eshel et al. [2015], Watabe-Uchida et al. [2017], O’Doherty et al. [2017]). Also, with the availability of new technologies, animal research was able to relate the *law of effect* to neuronal activity patterns, as reported by Athalye et al. [2018]: “Seeking evidence for a neural law of effect, we found that mice learn to reenter more frequently motor cortical activity patterns that trigger optogenetic VTA self-stimulation”.

Despite its remarkable success in explaining many experiments [Glimcher, 2011], the *Reward Prediction Error Hypothesis of Dopamine* has not been unquestioned and remains a topic of debate [Berridge, 2007, Dayan and Niv, 2008, O’Doherty, 2012, Beeler, 2012]. The role of dopamine in particular, has been studied extensively and research has put forward a much more complex role [Schultz, 2016]. It has been found that dopamine is involved in signaling the saliency of stimuli and the discovery of new actions [Redgrave and Gurney, 2006], the value of states and motivational vigor [Hamid et al., 2015], the selection of correct actions [Syed et al., 2015], or in signaling new, unconditioned stimuli (novelty) [Horvitz et al., 1997, Menegas et al., 2017]. Or, as Salamone and Correa [2012] summarize it:

> Traditional ideas about DA [Dopamine] as a mediator of “hedonia,” and the tendency to equate DA transmission with “reward” (and “reward” with “hedonia”) is

> giving way to an emphasis on dopaminergic involvement in specific aspects of motivation and learning-related processes, including behavioral activation, exertion of effort, cue instigated approach, event prediction, and Pavlovian processes.

Not only the role of dopamine has been challenged, but, more generally, also the neuronal circuits of value-based decision making remain a subject of debate. I conclude this summary with references to this still ongoing debate:

For example, the orbitofrontal cortex has been reported as being involved in a variety of cognitive processes related to learning and decision making, but many of them are questionned in [Stalnaker et al., 2015]. Also, Elber-Dorozko and Loewenstein [2018] conclude: “the claim that striatal neurons encode action-values must await new experiments and analyses”. Finally, even one of the main methods which has led to many insights about learning and the underlying neural circuits, parametric statistical analysis of fMRI, has been criticized for high false-positive rates [Eklund et al., 2016].

Despite this controversy, in the next section we describe a model that has been proposed as a neural implementation of model-free reinforcement learning.

1.3.2 A model of value-based learning with eligibility traces

As discussed before, a large body of research has established a close link between reinforcement learning theory and human learning and decision making. There is wide consensus that humans show hallmarks of value-based decision making [Glimcher and Fehr, 2013].

Decision making experiments have shown that humans learn the values of available actions and select actions by comparing their values (or utility) (see Rangel et al. [2008] for a review; but also Summerfield and Tsetsos [2015] for "irrational economic decisions"). The question then arises, where in the brain these values are represented and how comparison is performed. Many studies suggest that these computations are implemented in the cortico-basal ganglia - thalamo-cortical loop [Lisman, 2015]. A recent model for value-based learning and action selection, proposed by Collins and Frank [2014], is explained in Figure 1.4. In the cortico-basal ganglia - thalamo-cortical loop, the striatum is thought be involved in the representation and comparison of action-specific values [Samejima et al., 2005, Tai et al., 2012, Strait et al., 2015] The striatum receives reward prediction error signals via dopaminergic projections (see Fig. 1.5), facilitating learning of the action-specific values. When the reward is delayed, eligibility traces are necessary to bridge the temporal delay between the neural activity during action selection and the dopaminergic teaching signal. At the synaptic level, eligibility traces are linked to the *synaptic tagging and capture hypothesis* [Frey et al., 1997, Redondo and Morris, 2011, Gerstner et al., 2018]: Pre- and postsynaptic co-activity "tags" the synapse, while an additional factor (here, the dopamine mediated reward prediction error) is needed to consolidate the synaptic weight change as we saw in section 1.2.5. Experimental evidence for such eligibility traces at medium spiny neurons (MSN) in the striatum was provided by Yagishita et al. [2014] (in vitro, mice) and more recently by Fisher et al. [2017] (in vivo, rats). In both studies, the authors found a spine enlargement at MSNs to take place only if dopamine arrives during a critical time-window after excitatory input (0.3 to 2 seconds fin the first and 2 seconds in the second study).

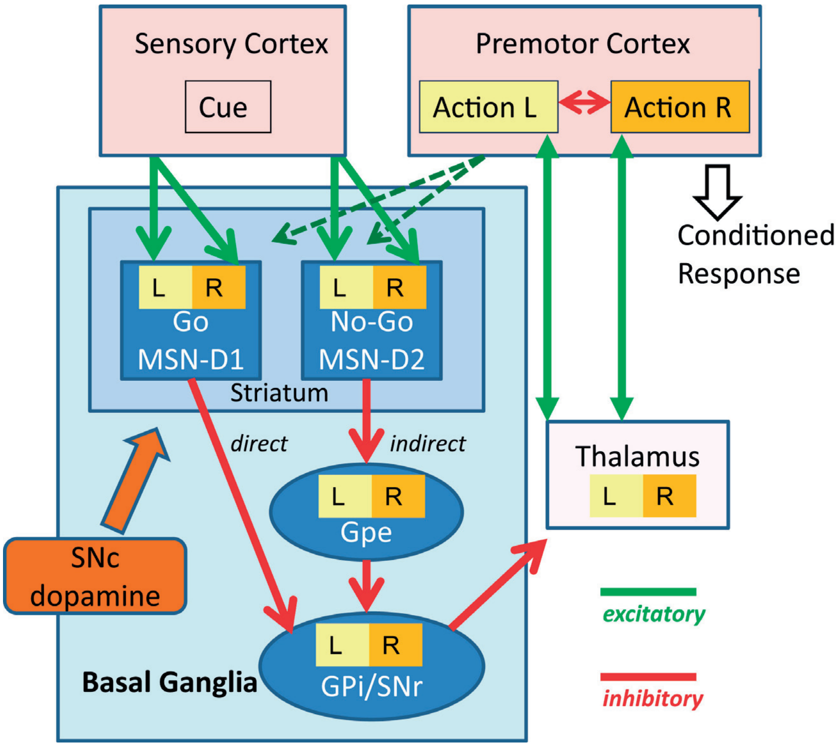


Fig. 1.4. Model for Action Selection and Learning in the "Go" and "NoGo" path.
Action Selection: Selection of conditioned actions is thought of as a fast, parallel comparison process implemented in the Cortico - Basal Ganglia - Thalamo - Cortical Loop [Collins and Frank, 2014, Lisman, 2015]. The sensory cortex signals the presence of a cue to the striatum. Two example candidate actions, left (L) and right (R), are processed in parallel streams, projecting from premotor cortex to different MSNs in the striatum. GPi/SNr are spontaneously active, inhibiting the thalamus and thereby reducing cortical activity. On the other hand, if GPi/SNr are inhibited themselves, the thalamus enhances the cortical activity of the corresponding candidate actions. This mechanism of inhibition or disinhibition is implemented through two different pathways, respectively: If a candidate action has been associated (through past experience) with a positive outcome, then the "Go" path is more active and inhibits the GPi/SNr through direct projections. On the other hand, if a candidate action was associated with negative outcomes, then the "NoGo" path dominates, and the GPi/SNr is *disinhibited* through the indirect pathway. Tonic dopamine can further excite (through D1 receptors) or inhibit (through D2 receptors) the "Go" and "NoGo" path. The final decision about which action to take is made in the premotor cortex.
Learning: If a chosen action is followed by a positive reward, the Go-pathway is reinforced by a process depending on the D1 dopamine receptors. On the other hand, a negative reward yields a reinforcement of the NoGo-pathway through processes related to the D2 receptors [Collins and Frank, 2014].
MSN: *medium spiny neuron*; D1, D2: *type D1 and D2 dopamine receptors*; Gpe: *globus pallidus externa*; GPi: *globus pallidus interna*; SNr: *substantia nigra pars reticulata*; SNc: *substantia nigra pars compacta*; Figure adapted from [Lisman, 2015]

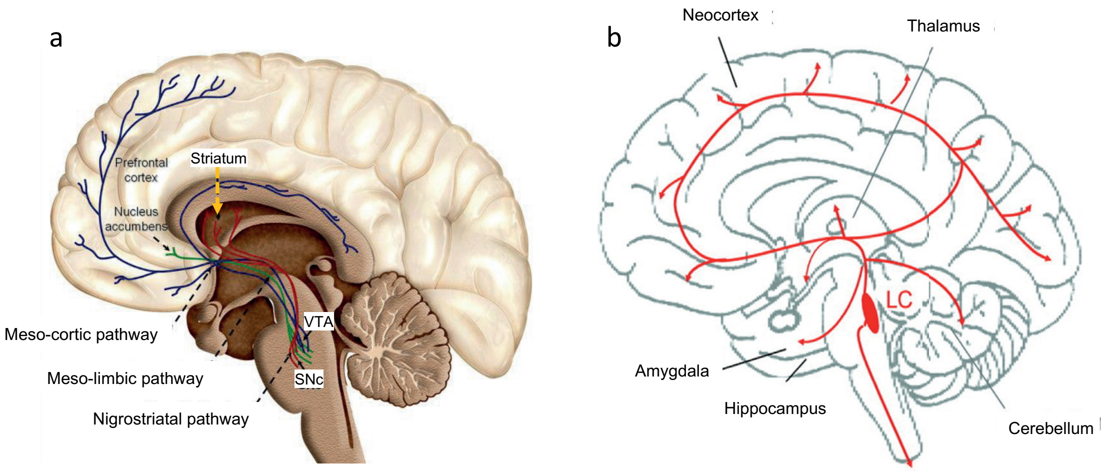


Fig. 1.5. Dopamine and Norepinephrine Projections. [**a**] Dopamine projections from the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc). Figure adapted from [Arias-Carrián et al., 2010]. [**b**] Norepinephrine projections from the locus coeruleus (LC). Figure adapted from [Benarroch, 2009].

1.3.3 Multiple learning systems and the tole of one-shot learning

In addition to model-free reinforcement learning, human behavior also shows hallmarks of model-based learning [Gläscher et al., 2010, Daw et al., 2011, Wunderlich et al., 2012, Doll et al., 2015a, O'Doherty et al., 2017]. That is, humans make use of observed state transitions to learn a model of the environment and use the model to infer the value of available actions. Figure 1.6 shows the abstract elements of model-based and model-free learning and decision making.

These two reinforcement learning approaches accumulate experience over many trials. This averaging over trials allows the algorithms to achieve good performance in stochastic environments, but makes learning slow. Recently, a third controller has gained attention: Episodic control [Lengyel and Dayan, 2008, Doll et al., 2015b, Chersi and Burgess, 2015, Gershman et al., 2017, Bornstein and Norman, 2017]. An episodic controller memorizes, without averaging, rewarded sequences of states and actions and repeats [5] the behavior when the same or similar situations are encountered (Fig. 1.6). This controller is not aiming at approaching an asymptotically optimal behavior. Instead, its goal is to make informed decisions already during early stages of learning [Lengyel and Dayan, 2008]. Memorizing a single, rewarded experience (one-shot learning) is sufficient (although not optimal in general) to determine future behavior.

Neurally, this controller has been associated with episodic memory in the hippocampus [Lengyel and Dayan, 2008, Gershman et al., 2017]. Humans can learn object classes from single examples and even generalize from it to previously unseen examples [Marblestone et al., 2016, Salakhutdinov and Tenenbaum, 2012]. Furthermore, recent learning and decision making experiments with humans, showed dissociable effects of incremental learning and one-shot learning [Rouhani et al., 2018, Duncan and Shohamy, 2016, Wimmer et al., 2014]. But these experiments were limited to state-reward pairings or one-step tasks where single actions were immediately rewarded.

Little is known about human one-shot learning in multi-step tasks. In the general discussion (chapter 3), we will review the results of our sequential decision making experiment, and the role of eligibility traces, in the light of one-shot learning.

---

[5] In the case of multiple stored experiences, averaging can still take place when the decision is taken. [Gershman et al., 2017]

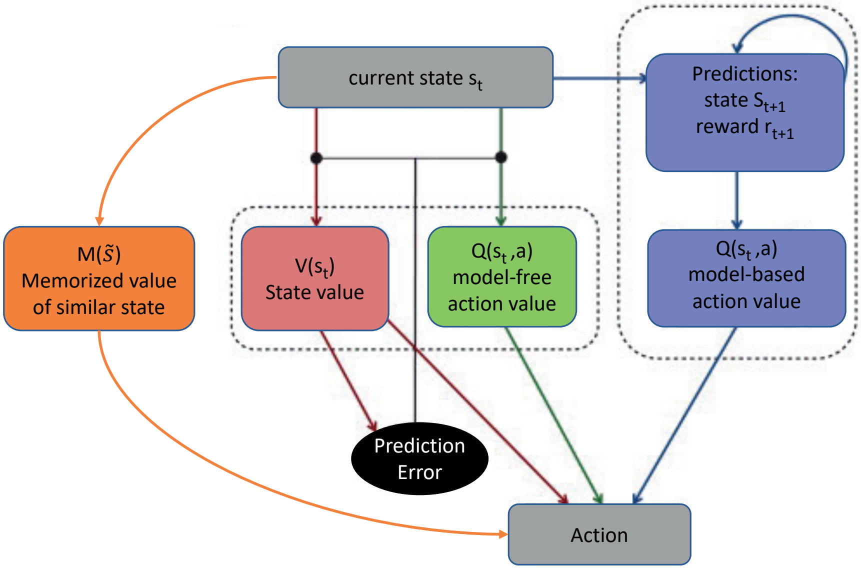

current state s_t
Predictions: state S_{t+1} reward r_{t+1}
$M(\tilde{s})$ Memorized value of similar state
$V(s_t)$ State value
$Q(s_t, a)$ model-free action value
$Q(s_t, a)$ model-based action value
Prediction Error
Action

Fig. 1.6. Multiple systems for learning and control. Different parallel systems to propose an appropriate action in the current state (gray boxes). Model-free systems base decisions on estimated state-values (red) or action-values (green). The Model-based system (blue) infers action-values from the predicted next states and rewards. A memory based controller (orange) stores single experienced states and outcomes. The value of an action is estimated from memorized outcomes of same or similar experiences. Figure adapted from [Daw and O'Doherty, 2013] and [Gershman et al., 2017]

1.4 Pupil dilation: a proxy for neural activity

In our experiments, we measure pupil dilation and relate it to the learning of the subjective value of states. In this section, I summarize literature describing the link between cognitive activities and pupil dilation. In short, the pupil dilation has been shown to correlate with various learning-related signals (e.g. reward prediction errors). These variations have been linked to neural activity in the locus coeruleus (LC), a nucleus that is the main source of the neuromodulator norepinephrine [Murphy et al., 2014, Joshi et al., 2016]. The projections from LC in the human brain are shown in Figure 1.5.

Pupillometry emerged as a tool to study mental activity in the 1960s [Hess and Polt, 1964, Kahneman and Beatty, 1966, Simpson and Hale, 1969] and has developed into a proxy for various brain states, including stress, pain, emotion, arousal, or cognitive load [Wang, 2011]. From an experimental perspective, pupil dilation has the advantage of being a relatively easy to measure continuous signal, providing non-invasive access to neural activity. Typically the pupil measurements are derived from high-speed video recordings: The eye tracking systems automatically extract gaze locations, eye-blinks and pupil dilation and provide a real-time stream of these measurements.

A light stimulus elicits a pupillary reflex, but the exact pupillary response depends not only on the stimulus itself but also on the cognitive state, as shown in Lowenstein et al. [1963]. The pupil response results from an interplay of the different parts of the autonomic nervous system: constriction is promoted by parasympathetic nervous activity, dilation by sympathetic activity [Wilhelm et al., 2001, Larsen and Waters, 2018]. Typically, increased sympathetic activity is accompanied by central inhibition of parasympathetic activity. In this regulation of autonomic activity and arousal,the LC-norepinephrine system has been found to play a central role [Samuels and Szabadi, 2008].

The connection between pupil dilation and neural activity in the LC has recently been further analyzed by Joshi et al. [2016] in a monkey study where they recorded pupil dilation and spiking activity in different brain areas. In several different conditions, the authors found a correlation between pupil dilation and neural activity in the LC. In one of their experiment, the unexpected presentations of auditory stimuli evoked a pupil dilation and LC spike rates. Analyzing the timing and activity patterns, the authors suggest that the two systems receive input from shared sources (rather than directly influencing each other), and that the common input might reflect an underlying change in attention and arousal.

Such a relationship between arousal and the LC-norepinephrine system has been suggested earlier in the context of *Adaptive Gain and Optimal Performance* [Aston-Jones and Cohen, 2005]. In particular, the study reports phasic LC activity in response to task-relevant events. Recordings from noradrenergic neurons in monkey LC showed selective, phasic response to conditioned stimulus (CS+) presentation [Aston-Jones et al., 1994].

Task-related pupillary responses were also observed in humans. In a behavioral experiment, recording fMRI and pupil dilation, O'Doherty et al. [2003] reported neural activity in ventral striatum and orbitofrontal cortex, which were compatible with reward prediction errors. Furthermore, the simultanously acquired pupillary signals, showed wider pupil dilations after the onset of conditioned stimuli (but before delivery of the predicted reward).

More recently, pupil dilation has been found to convey information about various signals [Laeng et al., 2012]. In a gambling task, where human participants were presented a card and had to predict whether a second card will be higher than the first one, Preuschoff et al. [2011] found the pupil to signal surprise. Other studies identified correlations between pupil dilation and recognition memory [Otero et al., 2011], attentional effort [Hoeks and Levelt, 1993, Alnaes et al., 2014, Unsworth and Robison, 2015], decision biases [de Gee et al., 2017, Eldar et al., 2018], decision uncertainty [Urai et al., 2017], decision timing [Einhäuser et al., 2010] , and encoding of memory [Kucewicz et al., 2018, Bergt et al., 2018].

We come back to the connection between pupil dilation and task-relevant variables in the experiments described in chapter 2.

1.5 Thesis contribution

In this thesis, I summarize and discuss the results obtained during my Ph.D. at the Laboratory of Computational Neuroscience (LCN) at EPFL. I worked under the co-supervision of Wulfram Gerstner, Professor at the School of Computer and Communication Sciences and Brain Mind Institute, School of Life Sciences, EPFL, and Kerstin Preuschoff, Professor at the Geneva Finance Research Institute, University of Geneva.

During my Ph.D., I collaborated in three interdisciplinary projects with Vasiliki Liakoni (LCN, EPFL) and He Xu (LPSY, EPFL). Individual contributions are clarified in the introduction of each project.

The overarching theme of this research was to investigate human learning and sequential decision making when reward is sparse and delayed. We designed and run different learning experiments with the aim of quantitatively linking behavior, reinforcement learning models, and (neuro-) physiological signals.

The main project of this thesis, presented in **chapter 2**, investigates the role of eligibility traces in human learning and sequential decision making. We designed a behavioral experiment which isolates the effects of eligibility traces. In behavior and pupil dilation we found signatures of learning with eligibility traces. We then discuss these results in the light of one-shot learning in the general discussion (chapter 3).
We conclude that state-action associations are formed from a single, rewarded experience. Moreover, reward does not only strengthen the state-action pair which received immediate reward, but also state-action pairs further back in the sequence.

In the second project, an fMRI study, we sought for neural correlates of different learning signals. A first goal was to develop a learning task which goes beyond the typically used two-step tasks, and to confirm previous findings (e.g. state and reward prediction error signatures) in a more complex environment. To this first part, Vasiliki Liakoni and me contributed equally. The second part of the project, led by Vasiliki Liakoni, focused on representations of states. Using multivariate pattern analysis, we investigated how representations of three state classes (start, goal, before goal) change over the course of learning. A summary of this project is given in **Appendix A**.

In the third project, led by He Xu, summarized in **Appendix B**, we studied human learning in a more complex environment, where only a single action at each state brings the participant closer to the goal, while all other actions lead back to *trap states*, far from the goal. As a consequence participants perform multiple iterations in the states far from goal, while only slowly discovering, state by state, the path to the goal. With this design, the discovery of a new state implicitly signals success. Modeling showed that standard reinforcement learning algorithms failed to explain the behavior in this task. Inspired by human behavior, we developed a hybrid model which combines both model-free and memory-based learners, mediated by an internal feedback signal related to the novelty of states.

2 Evidence for behavioral eligibility traces in human learning

Contributions

Designed the experiment: Marco Lehmann, Vasiliki Liakoni, Michael Herzog, Kerstin Preuschoff, Wulfram Gerstner
Implemented the experiment: Marco Lehmann, Vasiliki Liakoni
Run the pupillometry experiments: Marco Lehmann, Vasiliki Liakoni
Run the EEG experiments: He Xu, Marco Lehmann
Analyzed the behavioral and pupil data: Marco Lehmann
Analyzed the EEG data: He Xu
Discussed and interpreted the results: Marco Lehmann, He Xu, Vasiliki Liakoni, Michael Herzog, Kerstin Preuschoff, Wulfram Gerstner
Wrote the manuscript: Marco Lehmann, He Xu, Kerstin Preuschoff, Wulfram Gerstner

2.1 Abstract

In many daily tasks we usually 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. One approach suggested by reinforcement learning theory relies on eligibility traces in the form of a decaying memory of the state-action history. Here we asked whether humans indeed make use of eligibility traces in a multi-step decision making task, and designed a learning experiment which isolates the effects of eligibility traces in human learning. In three experimental conditions, using visual, acoustic, and spatial cues, we found that behavioral and pupillary responses at non-goal states exhibit experience dependent changes after a *single* reward, consistent with eligibility traces in reinforcement learning. Moreover, regression analysis revealed a robust temporal profile of the pupil response to reward prediction errors across the three different stimulus modalities, indicating a shared underlying neural process. Finally, EEG recordings showed a correlation between reward prediction error and event related potential amplitude, suggesting a processing of reward related signals at non-goal

states in frontal areas. Our results strengthen the link between reinforcement learning theory and experiments by providing direct signatures of eligibility traces, one of the key factors underlying fast learning.

2.2 Introduction

Reinforcement learning algorithms have successfully been applied to the study of human learning and decision making [Pessiglione et al., 2006, Schonberg et al., 2007, Gläscher et al., 2010, Daw et al., 2011, Niv et al., 2012, O'Doherty et al., 2017]. An important factor enabling fast learning in multi-step tasks is the eligibility trace [Sutton et al., 1988, Peng and Williams, 1996]. Eligibility traces capture the intuition that not only the most recent action contributes to an outcome but actions further back in the state-action history have also contributed and should be reinforced. While this concept is 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, Gerstner et al., 2018], it is unclear whether humans use an eligibility trace to improve performance in subsequent trials.

A difficulty in the study of eligibility traces in human learning is that relatively simple tasks, typically used in human studies, cannot exclude alternative learning strategies. For example, in an episodic two-step decision making experiment with rewards at the final states [Gläscher et al., 2010, Walsh and Anderson, 2011, Daw et al., 2011], a simple temporal difference (TD) learner with eligibility trace (e.g. *TD-λ*), would update the value of both, the start and the intermediate state, after the reward received at the end of the first episode [Sutton and Barto, 2018]. In contrast, a learner without eligibility trace (e.g. *TD-0*) updates the value of the start state only during the second episode. However, for such two-step sequences, during the third and all subsequent episodes, the behavioral data does not permit a clean dissociation between the two learning strategies. This is why eligibility trace contributions are typically statistically inferred from behavior through model selection [Bogacz et al., 2007, Gureckis and Love, 2009, Walsh and Anderson, 2011, Daw et al., 2011, Tartaglia et al., 2017], providing *indirect* evidence for eligibility traces. Due to this methodological difficulty, physiological correlates of eligibility traces during behavior are scant.

In this study we close this gap by introducing a novel experimental design to extract direct signatures of eligibility traces. We exploit that reinforcement learning with eligibility traces makes two testable predictions. First, the behavior at a state two actions away from a rewarded goal is reinforced after a *single* episode, observable as a selection bias in episode two (Fig. 2.1). Second, the value of that state is increased when the goal is reached the first time, yielding a reward prediction error when the state is seen again in episode two. We hypothesized that this reward prediction error translates to pupil responses, a known marker for cognitive activity [Kahneman and Beatty, 1966, Beatty, 1982, Joshi et al., 2016] and learning-related signals [Preuschoff et al., 2011, Browning et al., 2015, Foroughi et al., 2017]. We tested these predictions by implementing a hidden experimental manipulation of the state sequences in episodes one and two (see Methods and Fig. 2.1).

We recorded behavior, pupil dilation and EEG from at total of 37 human participants performing three different versions of the learning task, using spatial, acoustic and visual cues. We found strong effects of eligibility traces on behavior and pupil dilation after a *single* reward.

Furthermore, regression analysis showed a modulation of the pupil dilation by trial-by-trial reward prediction error at intermediate states and revealed a similar temporal profile of this response across three different stimulus modalities. In addition, the EEG signals at non-goal states (recorded from 22 participants performing the spatial-cue condition) showed a correlation between frontal event related potential (ERP) amplitude and reward prediction error. The EEG result is compatible with previous FRN studies [Holroyd and Coles, 2002, Walsh and Anderson, 2012] which used probabilistic one- or two step tasks. It indicates that the reward prediction errors are indeed processed in the frontal area of the brain in the time window 200 to 400ms after state onset. The ERP reflects neural prediction errors not just in probabilistic one- or two-step taks [Holroyd and Coles, 2002, Walsh and Anderson, 2012] but also at intermediate states in a deterministic multi-step task.

Taken together, our results provide insights into eligibility traces as a fundamental factor underlying the human capability of quick learning and adaptation.

2.3 Results

2.3.1 Task design and experimental conditions

Our learning environment consists of six states plus one goal state (Fig. 2.1). In each state, the participant choose one of two available actions, ’a’ or ’b’, and explored the environment until they discovered the goal. When the participant discovered the goal state, an episode ended and the next episode started in a different start state. The participants were instructed to reach the goal state as often as possible within a limited time of 12 minutes.

Different models of reward learning make quantitatively different predictions about the choice probabilities of a decision maker when learning this task. Over the course of learning, these differences vanish as an optimal solution is approached. Only during a few trials in the early phase of learning, models with eligibility traces make predictions which are qualitatively different from a model without eligibility traces. The design goal for this experiment was to move participants into these specific early trials, while giving them the illusion of freely exploring an unknown environment.

Contrary to the classic sequential decision making task, we used the following trick. In the first episode, all participants started in the state S (Fig. 2.1) and chose either action ’a’ or ’b’. However, independent of their choice and unknown to the participants, the action brought them always to state D2. Similarly, from D2, participants deterministically transitioned to D1 and from D1 to G. In episode 2, participants started from state Y and moved to D2 independently of their action. Episode 3 started in state X and participants were moved to D1. In half of the experiments, we swapped these two episodes and tested the transition X-D1 in episode 2 (Fig. 2.2).

The actions chosen by the participants determined the action-outcome binding for the remaining episodes in this environment. If for example, a participants has chosen action ’a’ in state D2 to transition to D1, ’a’ will bring this observer to D1 also in future episodes. The action ’b’ will bring the participant to state Z (Fig 2.1). Hence, there is no stochasticity in the environment. After episode 3, the participants started from the randomly chosen states S, D2, X, Y, Z.

After seven episodes, participant started over in a new environment where we replaced the images with different ones and reset the action-outcome binding. Participants were told that they were exploring a new environment. We used this manipulation to increase the number of episodes 1 and 2, which are the crucial ones for this study

Solving a new maze would become trivial if the participants would discover the underlying structure and manipulation. This was avoided by adding variations of the state sequences (see Methods). While participants performed the learning experiments, we recorded their behavior and the pupil dilation; in the *spatial* condition also EEG.

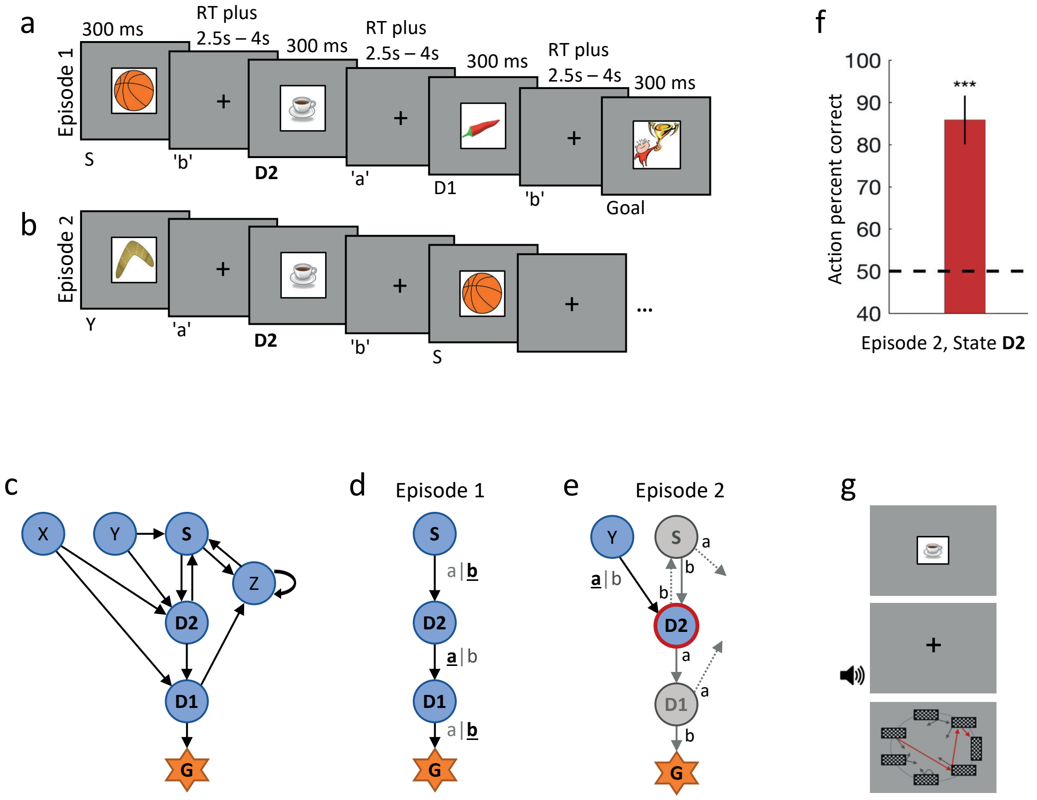


Fig. 2.1. Experimental design: **[a]** Episode 1. Participants reach the goal state G within three actions: 'b' at state S, 'a' at D2 and 'b' at D1. RT: reaction time. **[b]** Episode 2. The participant chose 'a' at Y and then saw D2 the first time after the rewarded goal state. While most participants chose the correct action (moving towards the goal) at D2, the participant in this example did not make the correct decision at D2 but chose 'b', moving away from the goal to S. **[c]** Structure of the environment. State transitions are predefined, but not the actions. **[d]** Episode 1. Unknown to the participants and independent of their actions, we guide each participant through the same states 'S' (Start), 'D2' (2 steps before goal), 'D1' (1 step before goal) to 'G' (Goal). Underlined actions, (S, 'b'), (D2, 'a') and (D1,'b'), show the decisions made by this particular participant (same as in [a]). To reinforce (D2, 'a') in episode 1, an eligibility trace is necessary to bridge the delay until G. **[e]** Epsdode 2 (corresponding to [b]). We move each participant from state Y to state D2 and test whether they repeat the decision made in episode one ('a' in this example). Solid arrows indicate the transitions defined through the participants behavior in episode 1; dashed arrows show the alternative action. **[f]** The fraction of correct actions at state D2 in episode 2, averaged across all participants, is above chance level 50%, showing a reinforcement from the single reward. **[g]** The same environment structure is implemented in three experimental conditions: *Top*: *Clip-art* condition. Each state is identified by a unique clip-art image, shown for 300ms. *Middle*: *Sound* condition: States are represented by unique sounds of 300ms to 600ms duration. *Bottom*: *Spatial* condition. States are identified by the location of a checkerboard. The goal state has a different orientation. Red arrows illustrate an example sequence S, D2, D1, G; participants see one rectangle at a time, flashed for 100ms.

To make our conclusions independent of a specific stimulus modality, we implemented three different versions of this task: in the (*spatial*) condition, states were identified by the location of a rectangle on the screen, in the *sound* condition by different acoustic cues, and in the *clip-art* condition by a picture in the center of the screen (Fig. 2.1 and Methods).

By building the transition graph as described, we created a learning condition with free binary choices, while still making sure that every participant moved through a controlled sequence of states, namely S - D2 - D1 - G, in the first episode. In episode two, we moved them into the critical state D2 where learning models with eligibility traces make qualitatively different predictions which we then tested.

2.3.2 Evidence for eligibility traces: Delayed reward reinforces state-action associations

Learning with eligibility traces predicts the reinforcement of state-action associations from a *delayed* reward. We investigated whether this prediction is observable as an action selection bias. Specifically, a reinforcement of the action taken in D2 in episode one, would require an eligibility trace, spanning the two-step delay from D2, across D1, to the goal state. In that case, we would expect a bias towards the reinforced action when the participants are in state D2 in episode two.

Each action is considered either correct (moving towards the goal) or wrong (moving away from the goal). We found that in the second episode, the percent-correct score in state D2, averaged across participants, is above chance level in all three experimental conditions (*spatial*: 62% correct ($p<0.05$), *sound*: 71% ($p<0.05$), *clip-art*: 85% ($p<0.001$), see Fig. 2.2). Evaluating the percent-correct scores in subsequent episodes showed that participants further improved their performance (Fig. 2.2).

Our results show a reinforcement not only of the action immediately followed by reward (state D1) but also of the decision made two steps before the goal state (state D2). This is consistent with predictions of reinforcement learning theories that include eligibility traces, but is incompatible with simpler algorithms.

2.3.3 Pupil dilation changes at states which received delayed reward

Next, we analyzed the pupillometric recordings for qualitative signatures of eligibility traces. The reward at the end of episode one potentially affects the subjective value of all states visited in episode one, namely S-D2-D1. Models without eligibility trace, predict a change of value only for the state immediately followed by reward (state D1), while eligibility traces affect all visited states, notably D2. We hypothesized that this change in value translates to a change in pupil response, because pupil dilation is a known marker for learning-related signals [Preuschoff et al., 2011, Lavín et al., 2014, Browning et al., 2015, Foroughi et al., 2017].

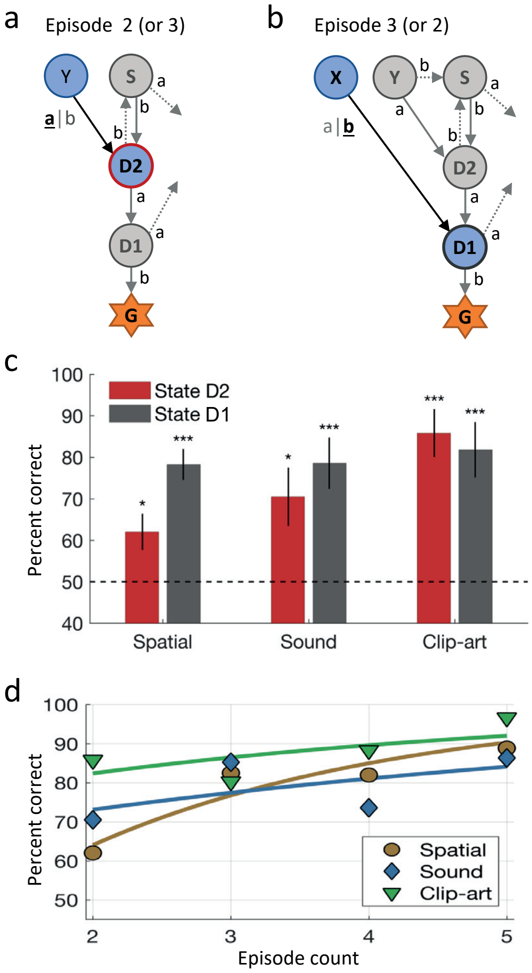


Fig. 2.2. Delayed reward reinforces state-action associations: The effect of the reward at the end of episode one, on the state-action pairs (D2, ·) and (D1, ·), is quantified in episode two. In half of the environments participants observe the transition Y-D2 in episode two. In the other half, X-D1 is experienced in episode two (and Y-D2 in episode three). [**a**] Participants move from a neutral start state Y to the state of interest D2. Reinforcement learning with eligibility traces predicts a bias towards the correct action ('a' in this example) at state D2. [**b**] In half of the environments, episode two starts at X and participants move to D1. [**c**] Percent-correct score of the decision at states D1 and D2, in episode two, averaged over all participants. In all three experimental conditions and at both states, D1 and D2, the bias towards the correct action is significantly above chance level (dashed horizontal line). Stars indicate significance levels: $^{*}p < 0.05$, $^{***}p < 0.001$ [**d**] Percent-correct at state D2 in later episodes. The episode count indicates the n^{th} episode in which state D2 was actually visited. Data points at Episode-count 2 are the same as in the left panel. To the scores (symbols) we fitted a saturating exponential learning curve.

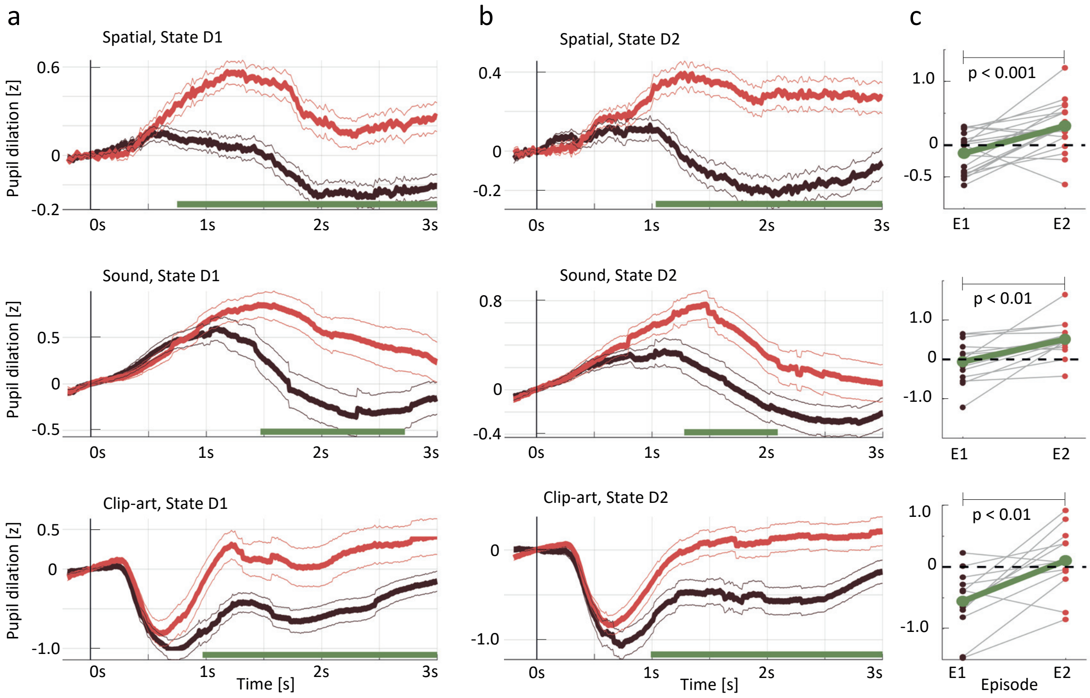


Fig. 2.3. Immediate and delayed reward affect pupil dilation in subsequent state visits. Pupil dilation response from 200 ms before to 3000 ms after stimulus-onset ($t = 0$), averaged across all participants. [**a**] Data for states D1. [**b**] Data for D2. *Top row*: *spatial*, *middle row*: *sound*, *bottom row*: *clip-art* condition. Black curves show the pupil diameter in episode one (before the participant has reached the goal state). Red curves show the pupil response at the same state in episode two, that is, after a single reward. Green lines indicate the time interval during which the two curves differ significantly ($p < FDR_\alpha = 0.05$). Thin lines indicate the pupil signal ±SEM. In both states and all three conditions, we observe a significant increase in the late pupil response in episode two compared to episode one. [**c**] Participant-specific mean pupil dilation at state D2 in $[1000ms, 2500ms]$ before (black dot) and after (red dot) the first reward. Grey lines connect means of the same participant. Green lines indicate a significant (paired t-test, p-values indicated in the Figure) increase of the means.

We extracted the time-series of the pupil diameter in the interval [0s, 3s] after the onset of state D2 (two steps from goal) and D1 during the first and second episodes (Fig. 2.1). The pupil trace at each state during the first episode, reflects the neutral response before having received a reward (Fig. 2.3, black traces). The red curves show the averaged pupil traces for the same states in episode two, when the participant went from Y to D2 (or from X to D1, respectively). All recordings show a significant change in pupil dilation in episode two compared to episode one. The comparisons were done per time point (paired samples t-test) and the significance levels were adjusted to control the false discovery rate (FDR, Benjamini and Hochberg [1995]). The difference reaches significance ($FDR_\alpha = 0.05$) at a time t_{min}, which depends on the condition and the state: *clip-art D1:* $t_{min} = 970$ ms (12, 39, 19); *clip-art D2:* $t_{min} = 980$ ms (12, 45, 41); *sound D1:* $t_{min} = 1470$ ms (15, 34, 19); *sound D2:* $t_{min} = 1280$ ms (15, 35, 33); *spatial D1:* $t_{min} = 730$ ms (22, 131, 85); *spatial D2:* $t_{min} = 1030$ ms (22, 137,130) (Values in brackets are: Nr. participants, Nr. pupil traces in episode one, Nr. traces in episode two).

Using clip-art images (bottom row in Fig. 2.3) the pupil trace shows an expected physiological contraction after the $300ms$ display of the clip-art images. In the *sound* condition, the significant separation of the curves occurs later than in the two other conditions and is of shorter duration. This might be related to the choice of the sound stimuli, which are of longer and variable durations ($300ms$ to $600ms$), and, according to participant's feedback, are more difficult to distinguish than the visual stimuli.

We also extracted the mean pupil dilation per participant. From each participant and each experiment, we obtained one pupil trace in episode 1 and one in episode 2. We then calculated a per-participant mean over experiments and over time (t in [$1000ms$, $2500ms$]) for episode 1 and for episode 2 (Fig. 2.3[c]). While the individual changes are noisy, the effect at the group level is significant (paired t-test, $p<0.001$ for the *spatial* condition, $p<0.01$ for the two others).

We performed a control experiment to exclude that the differences in the pupil traces are explained by the novelty of a state during episode one, and familiarity with the state in episode two, rather than by reward-based learning. A different set of participants saw a sequence of states (replays from the main experiments). They push a button in each state (freely choosing either 'a' or 'b'), and had to count the number of states from start to goal (to ensure that participants were focusing on the state sequence). Stimuli, timing and data analysis were the same as for the main experiments. In these control experiments, the strong difference after $1000\,ms$ in state D2, that we observed in Fig. 2.3[b], is absent (Fig. 2.6).

In summary, across three different stimulus modalities, the single reward received at the end of the first episode strongly influences the pupil response to the same stimulus in episode two. Importantly, this effect was observed not only in state D1 but also in state D2 (two steps before goal). Furthermore, simply observing repeated stimulus presentations, as in the control experiment, evoked a significantly different pupil trace. These results suggest that reward-based learning with eligibility traces are one of the driving forces of human pupillary responses.

2.3.4 Models with eligibility trace explain behavior better than alternative models

Having found direct evidence for eligibility traces in behavior and pupil dilation, we wondered whether behavior was better explained by reinforcement learning models which implement an eligibility trace than by alternative algorithms. We modeled eligibility traces $e_n(s,a)$ as a memory of all past state-action pairs (s,a) in an episode. At each discrete time step n, the eligibility of the current state-action pair is set to 1, while for all others it decays by a factor λ according to the following update rule [Singh and Sutton, 1996]

$$e_n(s,a) = \begin{cases} 1 & \text{if } s = s_n, a = a_n \\ \gamma\lambda e_{n-1}(s,a) & \text{otherwise.} \end{cases} \tag{2.1}$$

We initialize the table $e_n(s,a)$ at the beginning of each episode to 0. The parameter γ controls the speed of exponential discounting of a distal reward, commonly used in neuroeconomics [Glimcher and Fehr, 2013].

We considered seven common reinforcement learning algorithms as models for the behavioral data: *SARSA-λ* and *Q-λ* (Peng and Williams [1996], see Methods, Eq. 2.3) both implement an eligibility trace as defined in Eq. 2.1. *SARSA-0* [Rummery and Niranjan, 1994] and *Q-0* are their variants without eligibility traces (which is equivalent to setting the parameter λ in Eq. 2.1 to 0). As a third model for learning with eligibility traces, we considered, *Monte-Carlo Policy-Gradient* [Sutton and Barto, 2018]. This version of *Reinforce* [Williams, 1992] updates the action-selection probabilities only at the end of each episode and therefore requires the algorithm to keep a (non-decaying) memory of the entire state-action history of an episode. Alternative models of eligibility traces, more common in machine learning [Mnih et al., 2016, Sutton and Barto, 2018], use a memory buffer of *fixed* length n. For the task used here, we expect the performance of such *n-step* methods to be very close to the three models with eligibility traces, and therefore we did not add them to the analysis. We also considered a pure model-based algorithm, *Forward Learner*, which does not directly update state-action values (like *SARSA-λ* and *Q-λ*), nor action selection probabilities (like *Reinforce*). Instead, model-based algorithms use the observed transitions to estimate state-action-next-state associations and use this information to infer the actions which lead towards the goal state. Finally *Hybrid Learner* [Gläscher et al., 2010] combines the *Forward Learner* with the model-free *SARSA-0* into one algorithm. *Forward Learner* and *Hybrid Learner* do not rely on eligibility traces. As a null-model, we included a *Biased Random Agent*, which selects actions randomly with a left/right bias (fit to the data). We refer to (Sutton and Barto [2018], Gläscher et al. [2010] and Peng and Williams [1996]) for the pseudo-code and in-depth discussions of each algorithm.

To evaluate each algorithm's power in explaining human behavior, we fitted the free model parameters of each model to the behavioral data (separately for each experimental condition) and ranked the models using the Akaike Information Criterion (AIC, Akaike [1974]) (see Methods). Additionally we performed cross-validation to asses the robustness of the ranking and applied the Wilcoxon rank-sum test to evaluate significance (Table 2.1 and Methods).

We found *SARSA-λ*, *Reinforce*, and *Q-λ* to best explain the behavior in the conditions *clip-art*, *sound* and *spatial*, respectively. Importantly, they all ranked better than the best-fitting model without eligibility trace, which was the *Hybrid Learner* in all three conditions. Furthermore, while the individual models obtained different ranks across the conditions, when applied to the aggregated data, the algorithms with eligibility trace all significantly outperformed alternative models ($p < 0.001$).

2.3.5 Modeling quantifies an eligibility trace time-scale of 10 seconds

Next, we analyzed the fitted parameters to quantify the strength of the eligibility trace. Since the ranks of the three models with eligibility traces were not significantly different, we focused on one of these, *Q-λ*.

We found that the best fitting values (maximum likelihood, see 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 Fig. 2.4 for the uncertainties). While a value 0 would imply learning without eligibility traces, the values we find are all significantly larger than zero (p<0.001). The large values of λ indicate that participants strongly reinforce their decisions further back in the state-action history. However, directly interpreting the value of a single parameter can be misleading as λ is combined with other parameters that control action selection (cf. Methods and Eq. 2.4). Therefore, we evaluated the model's learning rules for the path S-D2-D1-G in episode one and thus obtained the action-selection bias when these states are seen in episode two. In agreement with the behavioral analysis, the model *Q-λ* (fitted to the data of *all* episodes) quantifies a strong bias in the action-selection probabilities after the first episode. At state D2, we calculate the following values for the different conditions: *clip-art*: 85%, *sound*: 77% and *spatial*: 73%. If we remove the eligibility trace (setting $\lambda = 0$) but keep only a left / right bias in the initialization of $Q(s, a)$, we find significantly lower values of 61%, 57% and 55%, respectively.

We then estimated the temporal aspects of the eligibility trace and considered a decay in continuous time rather than discrete steps. The inter stimulus interval ISI in our experiment is the sum of the reaction time of the participant and a random delay of 2.5s - 4s (Fig. 2.1). Therefore, an eligibility trace that affects actions two steps before the goal, needs to span a delay of at least two ISIs. We quantified this time-scale τ from behavior by replacing Eq. 2.1 with Eq. 2.5 and refitting the model parameters. We found maximum likelihood values for τ around 10 seconds (Fig 2.4). These results quantify an important role of eligibility traces in human reward learning and decision making.

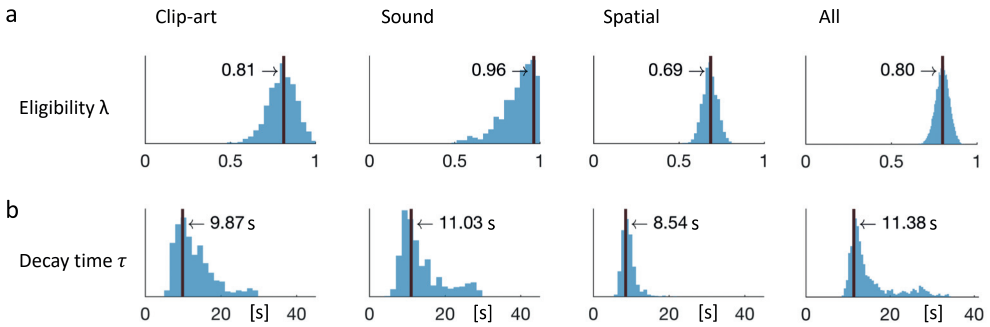


Fig. 2.4. Eligibility for reinforcement decays with a time-scale τ in the order of 10 seconds. The behavioral recordings of each experimental condition constrain the free parameters of the model Q-λ to the ranges indicated by the blue histograms (see Methods, Fig. 2.7, for all five parameters of Q-λ. **[a]** Distribution over the eligibility trace parameter λ (Eq. 2.1). Vertical black lines indicate the values that best explain the data (maximum likelihood). All values are significantly different from zero. **[b]** Modeling a time-dependent decay (Eq. 2.5) (instead of a discrete per-step decay with λ), the behavioral data constrains the time-scale parameter τ to around 10 seconds. Values in the column *All* are obtained by fitting λ and τ to the aggregated data of all conditions.

	Condition	Clip-art		Sound		Spatial		Aggregated
		AIC	Rank sum (k=7)	AIC	Rank sum (k=7)	AIC	Rank sum (k=11)	
	Model							all ranks
with elig. tr.	Q-λ	1234.8 $^{p(a)=.078}$	16	1489.1 $^{p(a)=.015}$	17	**6470.2 $^{p(a)=.001}$**	**20**	53
	Reinforce	1239.2 $^{p(a)=.109}$	17	**1486.8 $^{p(a)=.015}$**	**8**	6508.7 $^{p(a)=.015}$	30	55
	SARSA-λ	**1233.2 $^{p(a)=.015}$**	**13**	1495.2 $^{p(a)=.109}$	27	6502.4 $^{p(a)=.003}$	28	68 **p < .001**
without elig. trace	Hybrid	1271.3	26	1498.3	37	6536.6	52	115
	Q-0	1292.0 $^{p(b)=.015}$	44	1516.6 $^{p(b)= .046}$	34	6604.0 $^{p(b)=.003}$	51	129 $^{p(b) < .001}$
	SARSA-0	1289.5 $^{p(c)=.015}$	39	1518.2 $^{p(c)=.156}$	37	6643.3 $^{p(c)=.001}$	59	135 $^{p(c) < .001}$
	Forward Learner	1316.3	41	1500.6	36	6635.5	68	145
	Biased Random	1761.1 $^{p(d)=.015}$	56	1866.1 $^{p(d)=.015}$	56	7868.3 $^{p(d)=.001}$	88	200 $^{p(d) <.001}$

Table 2.1. Models with eligibility trace explain behavior significantly better than alternative models. For each experimental condition, the model with the lowest Akaike Information Criterion (AIC, evaluated on all participants performing the condition) is highlighted in bold. The significance of the differences between models is established using k-fold cross-validation, where each model is ranked k-times (see Methods). The sum of the ranks are shown in the column *rank sum*. The k pairs of individual ranks are used to compare pairs of models and obtain the indicated p-values (Wilcoxon rank-sum test). 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). P(b): the model *Q-0* compared with *Q-λ*. P(c): *SARSA-0* compared with *SARSA-λ*. P(d): *Biased Random* compared with the second last model (second largest AIC), which is *Forward Learner* in the clip-art condition and *SARSA-0* in the two others. In the **Aggregated** column, we compare the same pairs of models, taking into account all ranks across the three conditions. All algorithms with eligibility trace explain data better than algorithms without eligibility trace. Furthermore, differences among the three models with eligibility trace are not significant.

2.3.6 Reward prediction errors at non-goal states modulate pupil and EEG

We next asked if we can explain the changes in pupil dilation described in Fig. 2.3, as effects of the learning process. A key quantity that drives learning from reward, in models and humans, is the reward prediction error (RPE). While the RPE is often described as *'actual minus expected reward'*, one should note that RPEs can occur not just at rewarded (goal) states. Instead, by definition (Eq. 2.2), RPEs can arise at each time step from predicted (as opposed to actual) reward. For example, the start state of the second episode, Y, is a novel state cue and does not predict later reward, while the ensuing state D2 does (it has been associated with the reward during episode one through the eligibility trace). This difference yields a reward prediction error at the onset of state D2. We wondered whether such differences in *predicted* reward modulate the time course and amplitude of the pupil dilation.

By applying the model Q-λ to the behavioral data, we obtained the reward prediction error (RPE) associated with every state-on event. We then selected only the non-goal states and split them into two groups: one for which the calculated RPE is 0 and the other where the RPE is larger than the 80^{th} percentile. For each state-on event, we extracted the corresponding pupil signal and then compared the two groups. We found a strong modulation of the pupil dilation by RPE and a striking similarity of the temporal profile across the three experimental condition (Fig. 2.5).

Using regression analysis (see methods), we quantified these observations and found a significant parametric modulation of the pupil dilation by reward prediction errors at non-goal states (Figures 2.8 and 2.9). The extracted modulation profile reached a maximum at around $1 - 1.5s$ ($1300ms$ in the *clip-art*, $1100ms$ in the *sound* and $1400ms$ in the *spatial* condition) (Fig. 2.8), with a strong mean effect size (β_0 in Fig. 2.9) of 0.48 ($p < 0.001$), 0.41 ($p = 0.008$) and 0.35 ($p < 0.001$), respectively.

Thus, the pupil dilation is modulated by the reward prediction error calculated with a reinforcement learning model that implements eligibility traces. Furthermore, as we included only non-goal states in this regression, the pupil dilation reflects that reward value spills over to non-goal states. To further support this result, in particular the processing of reward prediction errors at non-goal states, we analyzed the EEG recordings (see Methods) and performed a regression analysis with the same reward prediction errors as used for the analysis of the pupil dilation. From the EEG signal, we extracted the mean amplitude of the event related potential (ERP) in the time window from 260ms to 405ms from four frontal electrodes (see Methods). We then calculated the linear regression between reward prediction error and ERP amplitude for each participant ($N = 22$), and found a positive slope for 18, a negative slope for 4 participants. The mean slope across participants was significantly positive (one-sample t-test, mean coefficient = 3.043, $t(21) = 2.75$, $p = 0.012$) (Fig. 2.10).

To summarize, the same reinforcement learning model with eligibility traces that enabled us to correlate the reward prediction error with pupil responses also allowed us to correlate the reward prediction error with the event related potential in prefrontal electrodes. The

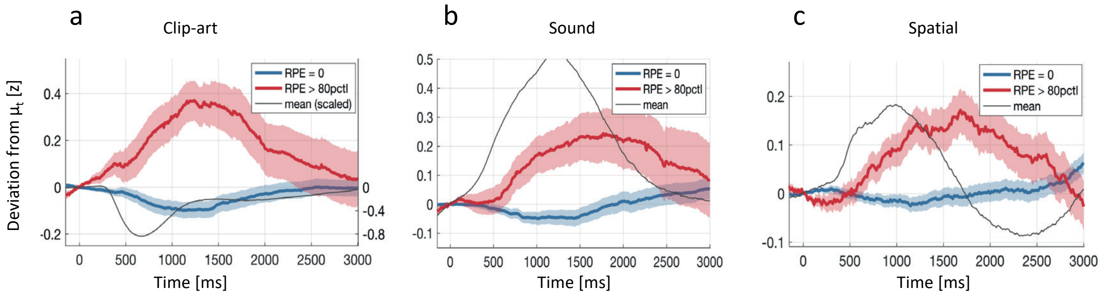


Fig. 2.5. Reward prediction error (RPE) at non-goal states modulates pupil dilation. Pupil traces were aligned at state onset ($t = 0ms$) and the mean pupil response μ_t (thin black curve) was subtracted. The deviation from the mean is shown for states with $RPE = 0$ (in blue) and for states with $RPE \geq 80^{th}$ percentile (red). Shaded areas indicate ± SEM. Only non-goal states are included in this analysis, and the pupil dilation therefore reflects the spreading of reward value to nonrewarded stimuli. Note that the mean pupil dilation (black) is different in each condition, whereas the reward related deviations from the mean have similar shapes. **[a]** Clip-art, **[b]** sound, and **[c]** spatial condition. In the clip-art condition, the mean pupil dilation (black line) is plotted in a different scale (right y-axis, values from -0.8 to 0).

regression analysis of the pupil and EEG data showed that the reward information transfers to non-rewarded states. Furthermore, the modulation of the two physiological signals suggests that humans process the reward prediction errors arising from learned values associated with non-goal states.

2.4 Discussion

In this study we examined whether humans make use of eligibility traces when learning a mulit-step decision making task. We found consistent evidence for learning with eligibility trace across three different experimental conditions in behavioral, pupillometric and EEG recordings.

First, we observed an important eligibility trace contribution to learning by analyzing the behavior during the second episode: in the state two steps before goal, the behavior showed a strong selection bias towards the correct action, implying a rapid reinforcement of the stimulus-response association after a single reward. Second, the pupillometric recordings at that state indicate a similar change after a single reward. Importantly, these two independent results do not lean on parameter estimation or summary statistics over many episodes, but are based on direct observations after a single reinforcement. Independent of the algorithmic details (be it *SARSA*, *Q*- or Value- learning), observing an effect at the penultimate state already in episode two excludes the possibility of a *TD-0*-like learning. Instead, it requires an eligibility trace to bridge the delay from the action, across another state-action pair, to the moment when the reward is finally obtained.

Third, we asked whether reinforcement learning algorithms that make use of eligibility traces explain the behavioral data better than alternative models without eligibility trace. We found *Q-*λ, *Reinforce* and *SARSA-*λ to have lower AICs (better explanation of the data) across all three conditions (*spatial*, *sound* and *clip-art*) than model-free or model-based algorithms without eligibility trace. The model-based algorithms learn state-action-outcome associations and thereby can be interpreted as models for prospective integration [Shohamy and Daw, 2015] or inference through associative links [Wimmer and Shohamy, 2012]. From our results we conclude that such model-based mechanisms are not dominating behavior during the early phase of learning.

We then analyzed the parameters of the algorithm *Q-*λ to quantify the eligibility trace. The best fitting values of λ are larger than 0 in all conditions and an algorithm $Q(0)$ without eligibility trace explains data significantly worse than *Q-*λ. Moreover, by evaluating the update rule of *Q-*λ after the first reward, we obtained the action selection probabilities in episode two at the state two steps before goal. We found the strongest eligibility trace effect in the *clip-art* condition, and the weakest in the *spatial* condition. This is in agreement with the behavioral analysis in which we calculated the percent-correct score directly from the actions taken in episode two only.

Alternatively to the per-state decay of the eligibility trace we also considered a time dependent decay and estimated a time-scale on the order of 10 seconds. Our experiment is not designed to decide whether the decay happens over continuous time or discrete events. Nevertheless the relatively slow decay reported here puts a constraint on the neural mechanisms implementing the memory needed to bridge the temporal gap between decisions and reward. Recent measurements of synaptic eligibility traces suggest a time scale of about $1s$ for dopamine-modulated plasticity in striatum [Yagishita et al., 2014, Fisher et al., 2017], about $2s$ for complex-spike plasticity in hippocampus [Bittner et al., 2017], and $2s$ or $5s$ for serotonin or norepinephrine modulated plasticity in cortex [He et al., 2015], but traces on the time scale of minutes have also been observed in hippocampus [Brzosko et al., 2017]. Moreover, modeling efforts over several decades have shown that reinforcement learning with eligibility traces can be implemented by neohebbian three-factor rules in computational models [Gerstner et al., 2018] that are consistent with these experiments.

Finally, EEG recordings confirmed the presence of a model-free reward prediction error at non-goal states. The event-related potential correlates significantly with trial-by-trial reward prediction error which was calculated using the Q-λ algorithm with eligibility trace. As the transitions of the task are deterministic, the prediction errors at non-goal states depend solely on learned state values, without being confounded with outcome probabilities. Furthermore, the time frame (from $260ms$ to $405ms$ after state onset) and location (frontal electrodes) are consistent with previous work [Holroyd and Coles, 2002, Walsh and Anderson, 2011], supporting the view that participants process a reward related signal at non-goal states.

Eligibility traces 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 direct evidence for eligibility traces in human learning from several different angles. The consistency and similarity of our findings across three experimental conditions suggests that the underlying cognitive processes are independent of the stimulus modality. It will be an interesting question for future research to actually identify the neural implementation of these memory traces.

2.5 Supplementary Materials and Methods

2.5.1 Experimental conditions

We implemented three different experimental conditions based on the same Markov Decision Process (MDP) of Fig. 2.1[a]. The conditions only differed in the way the states were presented to the participant.

In the first condition (*spatial*), each state was defined by the location (on an invisible circle) on the screen of a 100x260 pixels checkerboard image, flashed for 100ms, (Fig. 2.1[e]). The goal state was represented by the same rectangular checkerboard, but rotated by 90 degrees. The checkerboard had the same average visual intensity as the grey background screen. In order to collect enough samples from early trials, where the learning effects are strongest, participants did not perform one long experiment, but we broke the session into multiple independent experiments of seven episodes: After completing seven episodes, the experiment paused for 45 seconds, participants were instructed to close/relax their eyes. Then the experiment continued with a new environment: The transitions of the MDP were reset and the cues were rotated: states were represented by a 260x100 pixels checkerboard, the goal by a 100x260 pixels checkerboard.

In the second condition (*clip-art*), each state was represented by a unique 100 by 100 pixels clip-art image that appeared for $300ms$ 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 environments.

In the third condition (*sound*) each state was represented by a unique acoustic cue (tones and natural sounds) of $300ms$ to $600ms$ duration. At the goal state an applause was played. An experimental advantage of the *sound* condition is that a change in the pupil dilation can not stem from a luminance change but must be due to a task-specific condition. The location of the goal state rectangle in the spatial condition, as well as the assignment of clip-art and sound cues to the abstract states (S, D1, D2, X, Y,Z) was randomized in each environment.

The screen resolution was 1920x1080 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 × when participants had to enter their decision by pressing one of two push-buttons (one in the left and the other in the right hand). Subjects were instructed to reach the goal state as often as possible within a limited time (12 minutes). Prior to the actual learning task, they performed a few trials in each condition to ensure 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 $500Hz$, down-sampled to $100Hz$ for the analysis by averaging over 5 samples). From participants performing the *spatial* condition we recorded in parallel EEG and pupil data ($60Hz$ Tobii Pro tracker). An eye tracker calibration protocol was run for each participant. All experiments were implemented using the Psychophysics Toolbox [Brainard D. H., 1997].

The number of participants performing the task were: *sound* (SMI pupil): 15; *clip-art* (SMI pupil): 12; *spatial* (TET pupil & EEG): 22 participants; Control *sound* (SMI pupil): 7; Control *clip-art* (SMI pupil): 10; Control *spatial* (SMI pupil): 10. All participants were recruited from the EPFL students pool; all provided written, informed consent. The experiment was approved by the EPFL Human Research Ethics Committee.

2.5.2 Pupil data processing

Our data processing pipeline followed recommendations described by [Mathôt et al., 2017]. Eye blinks (including $100ms$ before, and $150ms$ after) were removed and short blocks without data (up to $500ms$) were linearly interpolated. In all experiments, participants were looking at a fixation cross which reduces artifactual pupil-size changes [Mathôt et al., 2017]. The time-series of the pupil diameter were extracted for each environment (= 7 episodes, see experimental conditions) and then normalized to zero-mean, unit variance per environment. This step renders the measurements comparable across participants and environments. We then extracted the pupil recordings at each state from $200ms$ before to $3000ms$ 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 (>220ms according to Mathôt et al. [2017]), but a symmetric baseline reduces small biases when different traces have different slopes around t=0ms. When analyzing these event-locked pupil responses, we excluded pupil traces with less than 50% eye-tracker data or with z-values outside $\pm 3\sigma$ in the window of interest.

2.5.3 Action assignment in the Markov Decision Process

Actions in the graph of Fig. 2.1 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 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 on from state D1 to Z, rather than to the Goal. This manipulation further implies that participants had to make decision against their potential left/right bias. Second, an additional state H (not shown in Fig. 2.1) was added in some environments. Subjects 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 one.

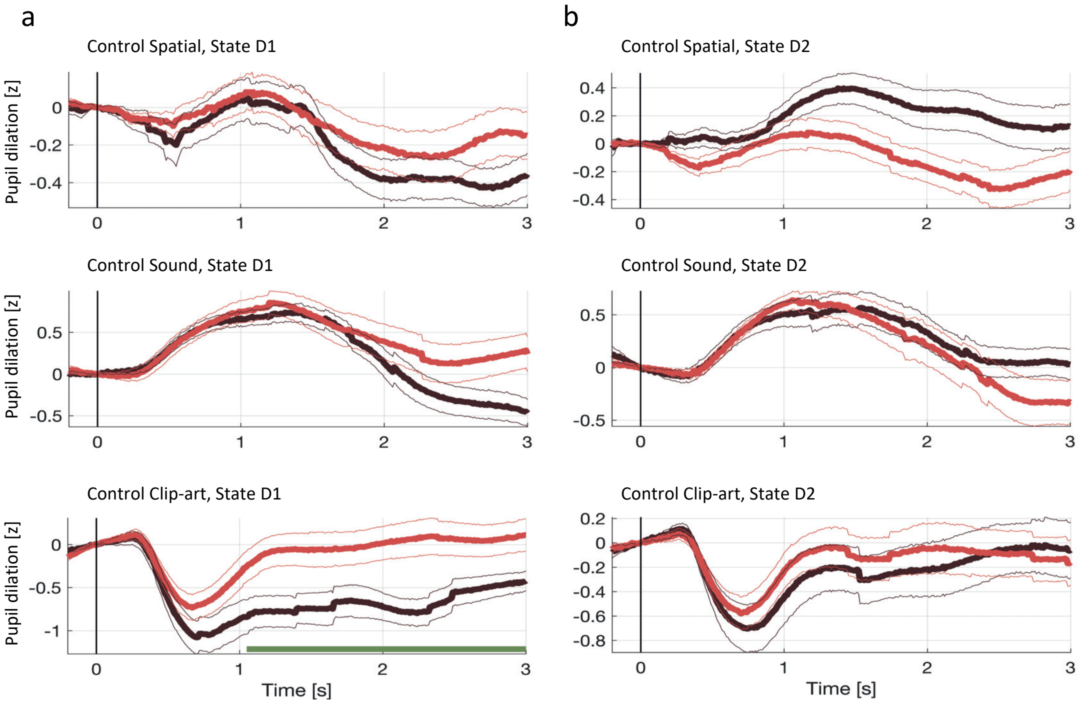


Fig. 2.6. Pupil dilation during control experiment. In the control experiment, different participants passively observed state sequences which were recorded during the main experiment. Data analysis was the same as for the main experiment. **[a]** Pupil time course after state onset ($t = 0$) of state 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 two is qualitatively different from the one in the main experiment (Fig. 2.3).

2.5.4 Reinforcement Learning Models

The algorithm *Q-λ* makes use of the eligibility traces $e_n(s,a)$ defined in the main text (Eq. 2.1), and the reward prediction error $RPE(n)$:

$$RPE(n) = r_{n+1} + \gamma \cdot max_{\tilde{a}}[Q(s_{n+1},\tilde{a})] - Q(s_n,a_n) \tag{2.2}$$

where r_{n+1} is the reward received after the transition from state s_n to s_{n+1} and $Q(s_n,a_n)$ is the expected future reward when taking action a_n in state s_n. This RPE is then used to update the state-action values $Q(s,a)$. Note that *all* Q-values are updated, but the individual changes are proportional to the eligibility of each state-action pair $e_n(s,a)$:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot RPE(n) \cdot e_n(s,a) \tag{2.3}$$

Where α is the learning rate. The table $Q(s,a)$ is usually initialized with zero, but here we identified for each participant the preferred action a_{pref} and initialized $Q(s,a_{pref})$ with a small bias, which is an additional free model parameter. The Q-values of Eq. 2.3 enter a softmax action selection with temperature T:

$$p(s,a) = \frac{exp(Q(s,a)/T)}{\sum_{\tilde{a}} exp(Q(s,\tilde{a})/T)} \tag{2.4}$$

As an alternative to the eligibility trace defined in Eq. 2.1, where the eligibility decays at each discrete time-step, we also modeled a decay in continuous time, defined as

$$e_t(s,a) = exp(-\frac{t - B(s,a)}{\tau})\ if\ t > B(s,a) \tag{2.5}$$

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 Eq. 2.2 is kept, while in Eq. 2.5 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 [Sutton and Barto, 2018], [Gläscher et al., 2010] and [Peng and Williams, 1996] for the pseudo-code and in-depth discussions of all other algorithms.

2.5.5 Parameter Fit and Model Selection

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 a left/right bias b.

To find the most likely values of those parameters, we pooled the behavioral recordings of all participants into one data set D. 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 as $P(D)$ is independent of θ^m, we can directly use the model likelihood $P(D|\theta^m)$:

$$P(\theta^m|D) = \frac{P(D|\theta^m)P(\theta^m)}{P(D)} \propto P(D|\theta^m) := f(\theta^m, D). \tag{2.6}$$

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's update rules (Eqs. 2.1, 2.2, and 2.3 in the case of $Q(\lambda)$) and the softmax function (Eq. 2.4) given a parameter sample θ^m. For computational reason, we use the log likelihood (LL). The sum is taken over all participants p, and all actions a_t a participant takes in each environment j:

$$LL(D|\theta^m) = \sum_{p=1}^{N} \sum_{j=1}^{E_p} \sum_{t=1}^{T_j} log(p(a_{p,j,t}|s_{p,j,t};\theta_m)) \tag{2.7}$$

For each model, we collected $100'000$ parameter samples (burn-in: 1500; collecting only every 10^{th} sample; 50 random start positions; proposal density: Gaussian with $\sigma = 0.004$ for parameters temperature τ and bias b, $\sigma = 0.008$ 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. Note that the parameter vector $\hat{\theta}^m$ could be found by a hill-climbing algorithm towards 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 each parameter i in θ^m (cf. Fig. 2.7). 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 comes 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 basic idea is the following: we repeated, K times, the same procedure as before (fitting, then ranking according to AIC) but now we used only a subset of participants

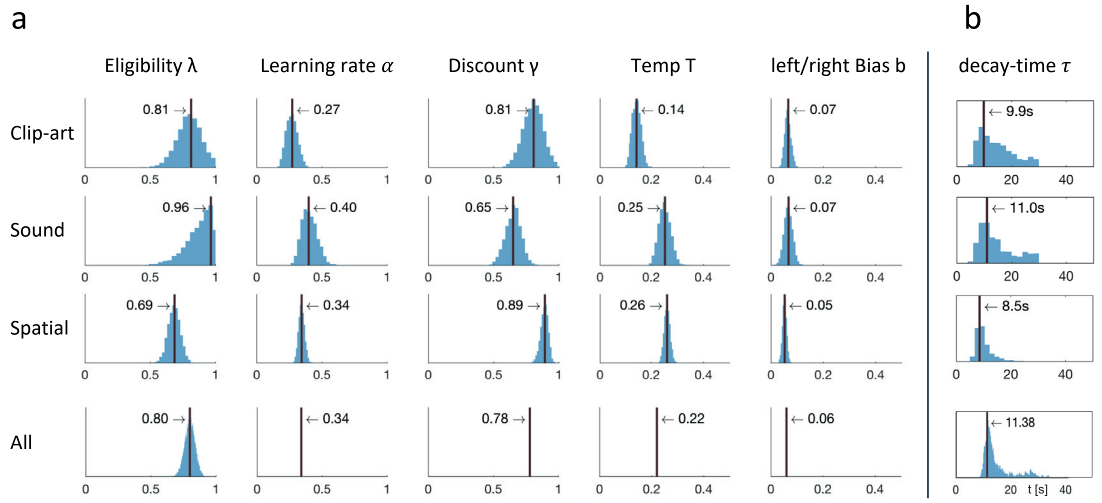


Fig. 2.7. Fitting results: behavioral data constrain the free parameters of *Q*-λ. **[a]** For each experimental condition a distribution over the five free parameters is estimated by sampling. The blue histograms show the approximate conditional posterior for each parameter (see methods). Vertical black lines indicate the values of the 5-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 2.5.

(the training set) to fit $\hat{\theta}_k^m$ and then calculated the LL_k^m and the AIC_k^m on the remaining participants (the 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 and evaluated each model on the same training/testing set in order to obtain comparable results. In each fold k, the algorithms 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 2.1). The cross-validation results were summarized by summing the K ranks (Table 2.1). The best rank sum a model could obtain is K, meaning it obtained the first rank in each of the K folds.

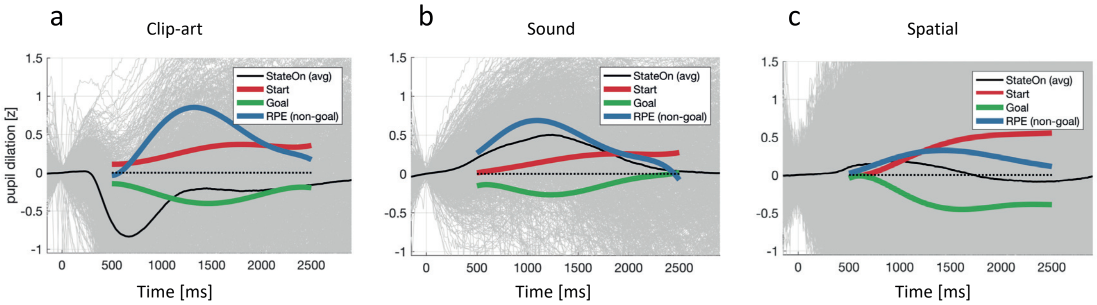


Fig. 2.8. Regression analysis reveals temporal profile of RPE modulation. Results for the conditions [**a**] clip-art, [**b**] sound and [**c**] spatial. Grey background curves show the preprocessed pupil recordings of individual trials aligned at stimulus onset ($t = 0ms$). The mean pupil response μ_t is plotted in black. As a visual guide, the dashed black line indicates effect size zero in the interval $[0.5s, 2.5s]$. All three regression results, "Start State" (in red), "Goal State" (in green) and reward prediction error at non-goal states (in blue), reveal a deviation from the mean pupil response. Although the mean curve μ_t (in black) is very different for the three conditions, the extracted event specific responses (in colors) are, up to some scaling, similar across conditions.

2.5.6 Regression Analysis

From Fig. 2.5 we expected a parametric modulation of the pupil dilation response by trial-by-trial RPEs. The aim of the regression analysis was to extract this unknown, event specific, pupil response (Fig. 2.8). Additionally we performed a non-parametric permutation test to evaluate the individual significance of each of the 5 Legendre components (Fig. 2.9).

The reward prediction errors (Eq. 2.2) used for the regression analysis (both, pupil and EEG) were 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 onset of the next action as in *SARSA*-λ), which corresponds well to the pupil traces locked to the state-on events. This can be seen from Eq. 2.2 where the value of the next state is evaluated: $V(s_{n+1}) := max_{\tilde{a}}(Q(s_{n+1}, \tilde{a}))$.

In our experiment, only the terminal goal state was rewarded; at non goal states, the reward prediction error depended solely on learned Q-values ($r_{n+1} = 0$ in Eq. 2.2), and at start states the reward prediction error 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 2.3 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 $k = 5$ (Fig. 2.8) as basis functions on the interval $-1 < t' < 1$. We use the indices p for participant and n for the n^{th} next-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

$$y_{p,n,t} = \mu_t + \sum_{k=0}^{5} RPE_{p,n} \times P_k(t') \times \beta_k + \epsilon_{p,n,t} \tag{2.8}$$

where the participant-independent parameters β_k were fitted to the experimental data (one independent analysis for each experimental condition). The models for "start state" and "goal 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.

In addition to the RPE modulation discussed in the main text, we also extracted the pupil dilation response at the start and goal states. We found a dilation of the pupil at the start of each episode, whereas at the goal state, where the reward prediction error was largest, we observed a *contraction* of the pupil. Interestingly, these temporal pupil profiles are quantitatively very similar across the three different stimulus modalities, suggesting a common neural process. It is known that changes in pupil dilation correlate with cognitive load [Kahneman and Beatty, 1966, Beatty, 1982], recognition memory [Otero et al., 2011], attentional effort [Alnaes et al., 2014, Unsworth and Robison, 2015], and even with encoding of memory [Kucewicz et al., 2018, Bergt et al., 2018]. We interpret the pupil traces at start and goal along those lines as markers for additional cognitive process beyond RPE-driven learning.

2.5.7 EEG Recordings and Analysis

From 22 participants performing the *spatial* condition, we additionally recorded EEG. EEG signals 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 EEGLAB toolbox [Delorme and 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 $35\mu V$ per sampling point were discarded. The baseline was removed using data 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 ERP analysis.

In order to identify the time window during which the event related potential (ERP) reflects the reward prediction error, we first focused on goal states. Reinforcement learning models such as $Q(\lambda)$ predict that the reward prediction error at the goal state decreases as the participant learns the task over multiple episodes. We extracted the ERPs from prefrontal electrodes when participants saw the goal for the first, third and fifth time in each environment, and found the ERP curves to reflect the trend $ERP_{Episode\,1}(t) < ERP_{Episode\,3}(t) < ERP_{Episode\,5}(t)$ in the time window $t \in [260ms, 405ms]$ (Fig.2.10[a]).

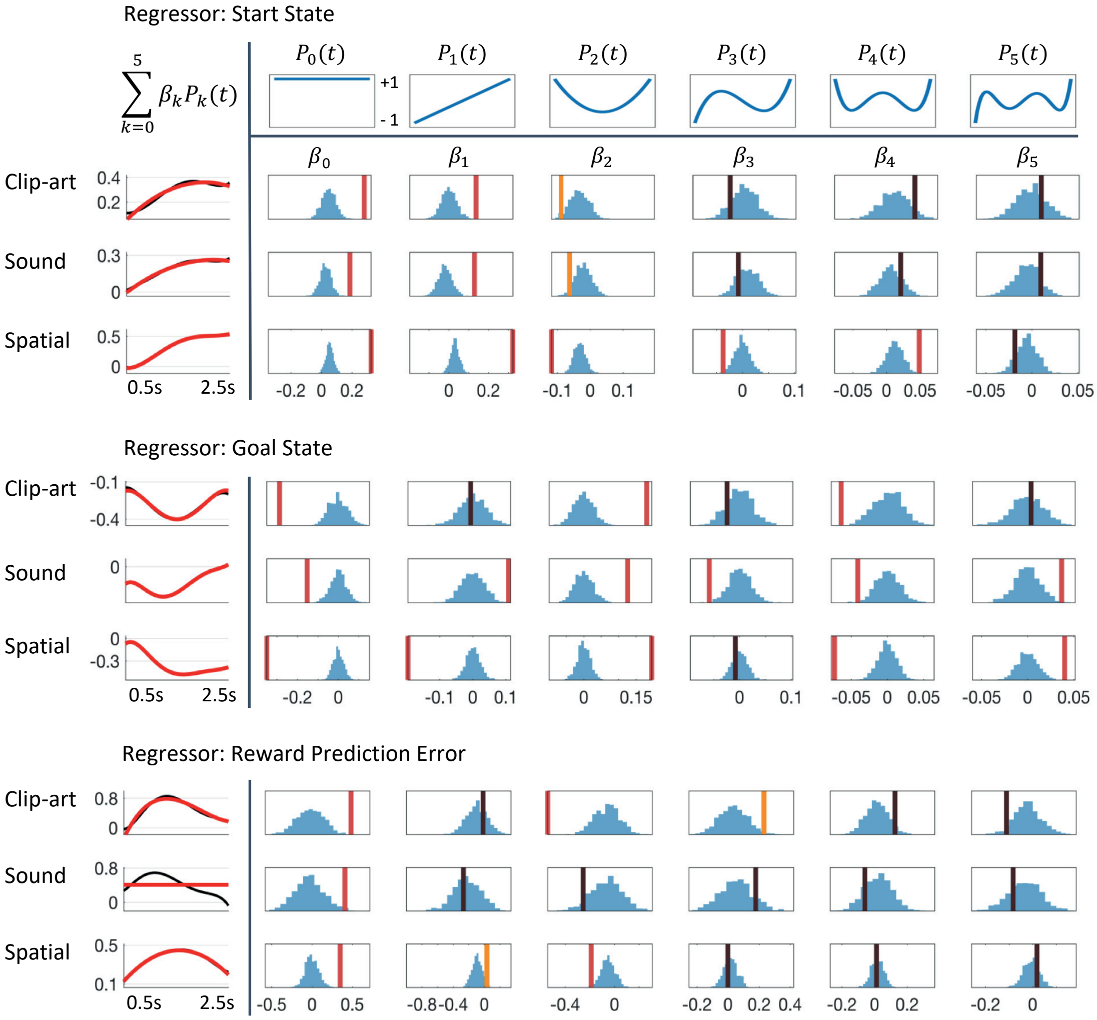


Fig. 2.9. Detailed results of regression analysis and permutation tests. The regressors are *top*: Start state event, *middle*: Goal state event and *bottom*: Reward Prediction Error. We extracted the time course of the pupil dilation in $[500ms, 2500ms]$ after state onset for each of the conditions, *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 are shown in the column below as vertical bars (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 6 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.

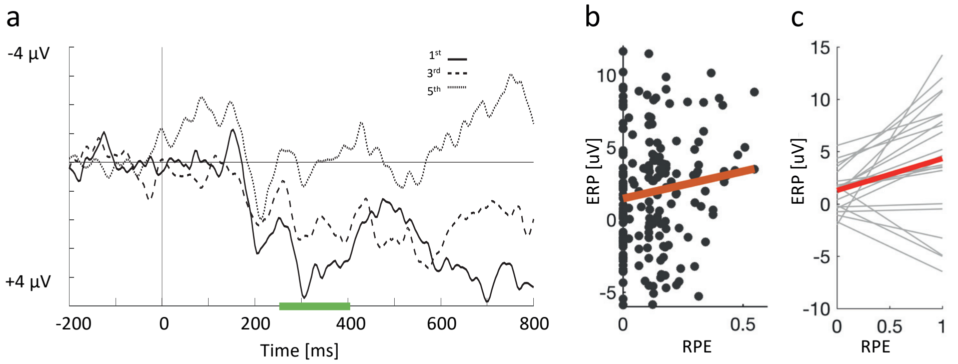


Fig. 2.10. EEG Data Analysis. [**a**] Time-locked ERP waves at the goal state (*spatial* condition, rectangle flashed from $t = 0ms$ to $t = 100ms$), averaged across all 22 participants. The three curves show the first, third and fifth reward. They significantly differ from $t = 260ms$ to $t = 405ms$, indicated by the green bar. This time frame is selected for the analysis in panel b and c. [**b**] Trial-by-trial mean ERP amplitude in t from $260ms$ to $405ms$ versus the corresponding reward prediction error for one participant. Only data from intermediate states (neither start, nor goal) is taken into account. The regression line is highlighted in color. [**c**] All regression lines from 22 participants. The red line shows the mean offset and mean slope. Slopes are different from 0 (one-sample t-test, mean coefficient = 3.043, $t(21) = 2.75$, $p = 0.012$).

2.5.8 Parameter fit and model selection: supplementary results

Fitting parameters per participant

The aggregated behavior of all participants constrains the free parameters to the ranges indicated in Figure 2.7. We wondered whether we can extract information about individual learning by fitting the parameters for each participant separately, and whether pupil dilation correlates with individual learning rates. It turned out that the noise level of the pupillary data of individual participants did not allow to draw conclusions.

Also, the behavioral data of single participants is not sufficient to constrain the parameters sufficiently. We addressed this problem by replacing the uniform prior in Eq. 2.6 with a multivariate Gaussian distribution. We used the parameter value that best explained (maximum likelihood) the aggregated data as the mean of the prior and $\sigma = 0.25$. Effectively, this prior shrinks the parameter space to a region of interest, acting as a regularizer. Results for two participants are shown in Fig. 2.11. We then compared the individual parameters and found that some parameters are strongly correlated, as indicated in Fig. 2.12.

Model Selection and Cross-Validation

When considering different RL algorithms as models of behavior, how does one identify the "best" model, and how does one quantify the difference to alternative models? It is common to base model selection on the Akaike Information Criterion (AIC) or the Bayesian Information

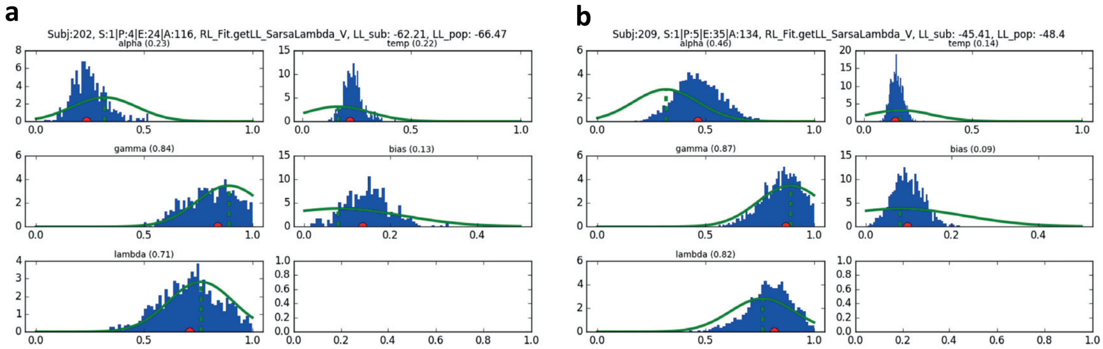


Fig. 2.11. Sarsa-λ parameter fit per participant The 5 dimensional parameter vector of Sarsa-λ was estimated individually for each participant by sampling (MCMC). The blue histogram summarizes the samples. The maximum likelihood value is marked in red. The green lines show the prior distribution over the parameters. For example, the learning rate alpha is centered at 0.32, which is the maximum likelihood value obtained from fitting the data of *all* participants. [**a**] and [**b**] are the fitting results for two different participant. Using the prior distribution constrains the parameter space and acts as a regularizer, while the data is still sufficiently informative to "push" the posterior away from the prior. This allows to identify groups of participants with learning rates "smaller" or "larger" than the average. For example, the participant in [a] has a lower learning rate (alpha=0.23) than the participant in [b] (alpha = 0.46).

Criterion (BIC). Here, we do not discuss the theoretical or philosophical support for either of the two criteria (which can be found in Burnham and Anderson [2004]). Instead, we apply a general rule of thumb and compare it with the results we obtained using cross-validation.

Burnham and Anderson [2004] gives the following rule to tell how much better one model is supported by the data than another: First, the model i with the lowest AIC is identified. Then the *difference* ΔAIC_k to an alternative model k is calculated and interpreted: Models with $\Delta AIC_k \leq 2$ have *substantial support*, and models for which $\Delta AIC_k \geq 10$ have *essentially no support*. Similar rules apply for the BIC. While serving as a rough guideline, this rule does not give a quantitative answer to the question whether differences are significant. Alternatively to the ΔAIC_k, *evidence ratios* or *Akaike weights* could be used [Burnham and Anderson, 2004], but the question of significance remains.

In section 2.5.5 we compared several algorithms. In table 2.1 we indicate both, the AIC scores and the significance based on cross validation. We note that using cross-validation and rank statistics, we obtain a more conservative interpretation of the data than the rule of thumb would give. For example, in the clip-art condition, we have AIC Q-λ = 1234.8 and AIC Hybrid = 1271.3. The rule of thumb indicates that the Hybrid model has "no support", because the difference in AIC is 36.5, whereas the statistical test on cross-validation results fail (p=0.078).

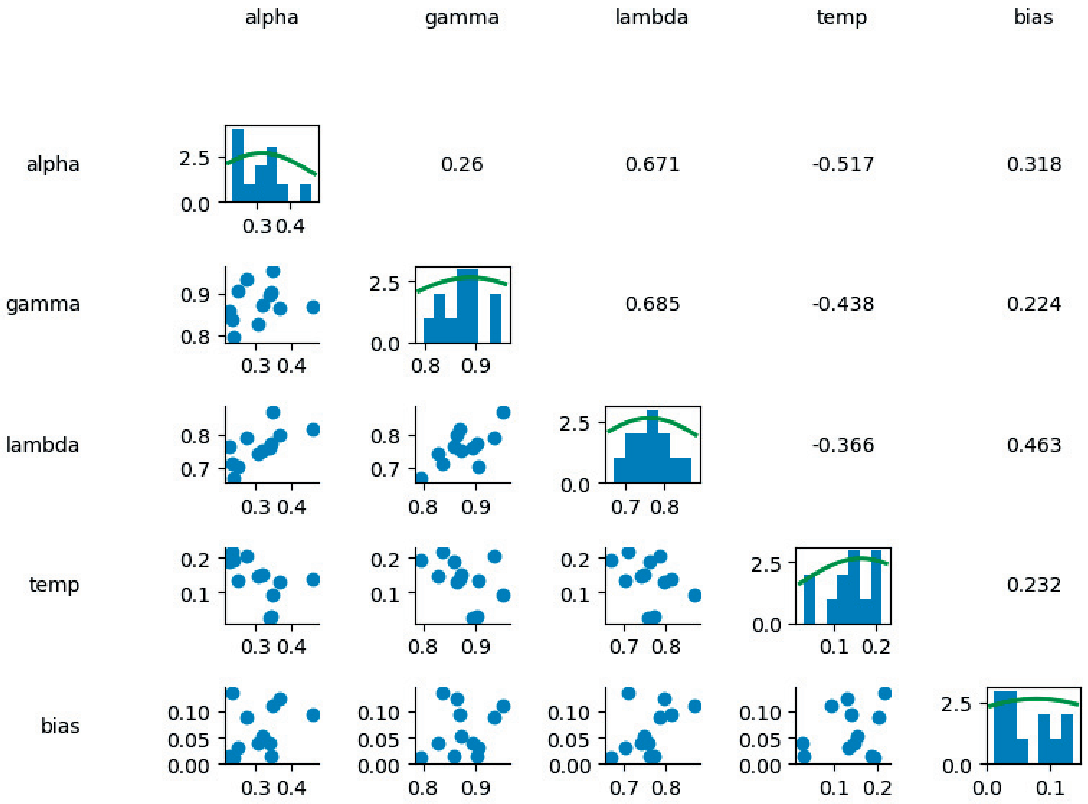


Fig. 2.12. Correlation of SARSA Parameters For each participant, a single parameter vector is fitted. Histograms in the diagonal show the distribution per parameter. Each scatterplot below the diagonal shows two joint parameters, dots represent participants. The corresponding pairwise correlations are shown above the diagonal.

3 Discussion

3.1 One-shot learning and eligibility traces

In the previous chapter we experimentally showed effects of a single reward occurrence on human decision making. We explained behavior in terms of model-free reinforcement learning with eligibility traces. However, recent experimental studies have shed light on memory encoding from single experiences, the so-called one-shot learning (section 1.3.3). Here I compare these experiments with our task and then discuss our results in the light of one-shot learning and episodic control.

3.1.1 One-shot learning and episodic control in human experiments

Wimmer et al. [2014] investigated the interaction between model-free reinforcement learning and one-shot learning in a single-step task. At each trial participants were shown two different images, placed within a green or a blue square. Participants had to choose one of the two options and received a time-varying, color-dependent reward. The images were unique in each trial and did not influence the reward. During a second part of the experiment (one day later) the participants were shown the images again and had to rate whether they had seen each image before or not. By this design, the authors could test behavioral effects of reward-based learning, as well as the memorization of items. The authors found the recent reward history to negatively covary with memory encoding and discussed this result as an effect of competition between model-free reward learning and episodic memory encoding. In contrast to our experiment, the memorization of the unique state images was not task relevant.

In a similar task, Rouhani et al. [2018] have shown one-shot learning of stimulus-reward associations. Participants had to estimate the average reward associated with categories (e.g. indoor vs. outdoor) of unique images. Later, the recognition of the images was tested. Interestingly, the authors found two distinct effects of reward prediction errors: they not only drive incremental value learning, but also support the one-shot creation of episodic memory.

Unlike our task, this experiment did not involve actions.

The influence of episodic memory on human decision-making was studied experimentally by Duncan and Shohamy [2016]. In three variations of a context-choice-reward experiment the authors showed that a single exposure to a stimulus, followed by reward, guided future decisions. This choice bias was also induced if the reward was delayed by 1.5 seconds. This design comes closer to our experiment as it has a stimulus-action-reward structure, but it is not a multi-step task.

The experiments discussed in this section suggest to consider two different learning systems: one-shot learning (and episodic control in the case where actions are involved), as well as incremental reinforcement learning. The conceptual difference between these experiments and ours is that we employ a task that involves the selection of actions over multiple steps. In the next section we discuss whether our results are compatible with one-shot learning.

3.1.2 Evidence for one-shot learning and the role of eligibility traces

In our task, effects of one-shot learning would be observable in episode two, when the states along the trajectory to the goal are visited for the first time after a single, rewarded sequence. We will now consider our results from the analysis of behavioral and pupillometry data separately.

First, as we saw in chapter 2.3.2, there is an action selection bias at the states situated one and two steps before goal (Fig. 2.2). Clearly, the correct action in episode two is the repetition of a single, rewarded action in episode one. The decision to repeat this single rewarded action would be compatible with one-shot learning. But whether this action has actually been chosen by an episodic controller or by a value-based system can not be answered by our experiment.

We now turn to our results on pupil dilation (Fig. 2.3.3). When we compared the pupil dilation responses in episode one to episode two, we attributed the dilation to the reward prediction error. An alternative interpretation is that the wider pupil dilation in episode two stems from the recognition of the state. Kucewicz et al. [2018] found stereotypical pupil dilation responses during recall and memorization of words. A wider pupil during recognition of previously seen items has also been reported by Otero et al. [2011]. These results suggest that the pupil is indeed a marker for memory processing and our pupillary data can be interpreted accordingly. In this view, the result of the control experiment (Fig. 2.6), where the widening of the pupil in episode two is absent, suggests that reward is necessary for memorization. This observation is compatible with the aforementioned results of Rouhani et al. [2018].

When considering reward driven one-shot learning over trajectories, i.e. sequences of actions, we also have to consider the role of the eligibility traces. The formation of episodic memory of states and actions which were followed by delayed reward, requires an eligibility trace for the same reason as explained for value learning (Section 1.3.2). In this case, the eligibility trace would have to support the formation of episodic memory in hippocampus. Such eligibility traces have been identified in the hippocampus of mice by Brzosko et al. [2015, 2017], Bittner et al. [2017].

Our experiments show effects of one-shot learning in human behavior and in pupil dilation responses. Eligibility traces support fast learning from delayed reward. In our view, these results are compatible not only with model-free control, but also with episodic control. Whether one-shot learning implements an update of values in a model-free controller, or the creation of memories in an episodic controller, or even both in parallel, cannot be concluded from our experiments and would be an interesting question for future research.

A Human learning and sequential decision making. An fMRI Study

Using fMRI, we investigated human learning during a sequential decision making task. A first goal of the study was to develop a learning task which goes beyond the typically used two-step tasks, and to confirm previous findings (e.g. state and reward prediction error signatures) in a more complex task.

The second part focuses on representations of states. Using multivariate pattern analysis, we investigate how representations of three state classes (start, goal, before goal) change over the curse of learning. This second part of the project is ongoing work, led by Vasiliki Liakoni.

Only the first part of the project is summarized here.

Contributions

Designed the experiment: Vasiliki Liakoni, Marco Lehmann, Kerstin Preuschoff
Implemented the experiment: Vasiliki Liakoni, Marco Lehmann
Run the fMRI experiments: Vasiliki Liakoni
Analyzed the behavioral data: Marco Lehmann, Vasiliki Liakoni
Analyzed the fMRI data: Vasiliki Liakoni
Discussed and interpreted the results: Vasiliki Liakoni, Marco Lehmann, Johanni Brea, Wulfram Gerstner, Kerstin Preuschoff
Wrote the manuscript: Vasiliki Liakoni, Marco Lehmann, Kerstin Preuschoff, Wulfram Gerstner

A.1 Abstract

Recent advances in computational, behavioral and cognitive neuroscience have indicated that humans employ multiple strategies to learn from the outcome of their actions. Nonetheless, our understanding of learning behavior is still largely restricted. Current experimental tasks are often simple compared to the real world and so are the reinforcement learning (RL) models used to explain behavior in these experiments.

Here we employ a novel multi-step sequential decision making task alongside a larger repertoire of algorithms to explain human learning behavior. To facilitate fMRI data analysis, the experiment is designed to de-correlate signals of different strategies. Twenty-three human subjects performed the task in an fMRI scanner. We considered the following algorithms: three model-free (MF) value-based algorithms , one model-based (MB) algorithm, an MF-MB hybrid learner and a policy gradient algorithm. We find correlates of MF prediction errors in the ventral striatum and other areas and MB correlates in the inferior frontal gyrus and insula.

Our results corroborate the existence of two systems in the brain performing MF and MB computations, in agreement with previous studies. Interestingly, our behavioral data are best explained by a policy gradient algorithm and by an update of actions based on eligibility traces and end-of-episode reward, rather than intermediate errors.

A.1.1 Introduction

Animals and humans tend to repeat rewarded actions and have reward prediction capabilities [Dayan and Yu, 2006]. Model-free (MF) RL algorithms which incrementally update state-action values or policies [Sutton and Barto, 2018] successfully explain many aspects of reward-based learning. Contrary to behavior though, MF algorithms are inflexible to changes in the environment or task structure. Model-based (MB) RL algorithms learn a model of the environment, i.e state transitions and can cope with such changes, at the cost of higher computational expenses [Gershman et al., 2017].

Human fMRI studies have found signatures of both MF and MB algorithms in separate, as well as overlapping brain areas. But which strategies best describe human behavior in which situations, as well as how strategies are implemented and combined in the brain are still open questions. On the experimental side, apart from technical limitations of recording and imaging procedures, research trying to address the above questions has mostly employed two-stage tasks (Daw et al. [2011], Gläscher et al. [2010]), where the temporal credit assignment problem is less pronounced and computational expenses of MB learning become minor. Attempts to scale up the task complexity in brain imaging experiments come together with the challenge of dissociating different learning signals, as different strategies tend to give correlated predictions.

Here, we employ a new task with an experimental manipulation that disentangles different learning signals. We corroborate previous fMRI findings in this multi-step scenario. Furthermore, the behavioral data is best explained an algorithm which updates policy parameters at the end of the episode. Such an update requires a memory of the past state-action pairs, akin to eligibility traces.

A.2 Methods

A.2.1 Experimental design to separate MF from MB prediction errors

The task is situated in a state-space with 7 circularly arranged fractal images (states) with 2 possible actions at each state (apart from the goal which is a terminal state) (Figure A.1[a]). At the beginning of each episode subjects are shown an initial state. Choosing an action results in a transition to another state. Subjects continue to choose further actions until they reach the goal, which completes an episode (Figure A.1[b]). Their task is to learn how to reach the goal in the smallest number of actions. The goal state is visually distinguishable from all other states and is indicated to the subjects beforehand. State transitions are deterministic, with occasional "surprise trials", explained further below.

During the experiment we perform an online evaluation of the MF SARSA(λ) algorithm [Sutton and Barto, 2018] and the MB Forward Learner (Gläscher et al. [2010]) based on the subject's choices. This gives us access to an online estimate of the reward prediction error (RPE) and the state prediction error (SPE) used in the two algorithms, respectively. The SARSA(λ) approximates the $Q(s,a)$ values, i.e the estimated expected reward starting from state s and action a, incrementally via the RPE . On the other hand, the Forward Learner estimates the transition probabilities via the SPE and uses them to compute the $Q(s,a)$ values via the Bellman equation. For the online RPE and SPE computations the choice of the parameters (e.g. learning rates α) was based on pilot experiments performed prior to this study. On a *surprise trial* subjects transit to a state s'' other than the one they have learned to expect as the outcome of action a in state s. If subjects expect to transition to state s' from s, they now transition to s'', chosen such that s' and s'' have similar $V(s)$ values ($V(s) = \max_a Q(s,a)$), namely $\left|V(s') - V(s'')\right| \leq \epsilon$, where ϵ is a small threshold. In this way, the experienced MF RPE will be low, but the MB SPE will be high, since the learned transition was violated (Figure A.1[c]). This novel experimental manipulation with online monitoring allows us to de-correlate the errors used in the learning algorithms.

At every trial the conditions for performing a surprise trial were checked online (i.e learned transitions and V values within the threshold). No surprise trials were employed within 3 trials since last surprise trial. Also, if there were 8 consecutive trials that no surprise trial has occurred, a randomly chosen unexpected transition was enforced in order to ensure some variability. Thus we have two types of surprise trials: (i) those that meet the threshold criterion on V-values and (ii) purely random transitions.

Each subject performed a 20-min block of the task in the fMRI scanner, which involved approximately 56 full episodes and approximately 235 state transitions per subject. From these approximately 15% were surprise trials. Prior to the experiment, subjects got familiar with the task outside and inside the scanner during short sessions of two episodes each, with different images and transitions than the ones used during the experiment. Furthermore, subjects were beforehand informed on the existence of surprise trials and on the fact that the

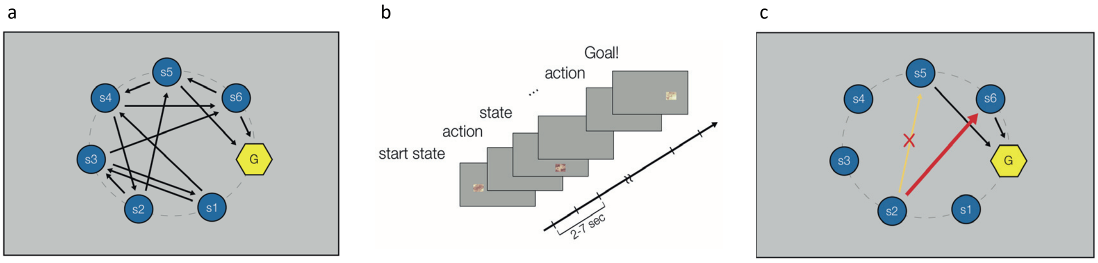


Fig. A.1. Sequential decision making task with surprise trials [a]: Seven states are positioned on an invisible circle. Arrows indicate the transition graph. Note that the transitions are deterministic and that the graph allows cycles. **[b]:** During the experiment, participants see a single fractal at the state specific location. After pressing a push-button, the image disappears and the screen stays grey for a few seconds (2-7 seconds random inter stimulus interval ISI). Subjects continue their decision making process until they find the goal state, which terminates one episode. **[c]:** While the transitions are deterministic most of the time, we introduce surprise trials in about 15% of all transitions. Here, the transition S2 to S5 (yellow arrow) is usually observed. But instead of moving the participant to the expected state S5, we move her to S6. This unexpected transition introduces a *high* state prediction error SPE. The new target state S6 was chosen such that the value is similar to S5, yielding a *low* reward prediction error RPE. This manipulation is crucial to decorrelate the two signals, RPE and SPE.

underlying transition matrix does not change. Fractal images, their locations on the screen and the actions' correspondence to left-right response buttons were randomised across subjects. We employed two different underlying transition matrices, also in a randomized way. Subjects were compensated with a fixed monetary amount for their participation, plus a small extra performance-based amount.

A.2.2 Reinforcement Learning algorithms

To explain subjects' behavior we considered multiple algorithms. As briefly mentioned above, the SARSA(λ) estimates the $Q(s,a)$ values with RPE-mediated updates. In the Forward Learner the model of the world, namely the transition probabilities from state s to s' when selecting action a, $T^{a}_{ss'} = P(s_{t+1} = s'|s_t = s, a_t = a)$ are estimated using the SPE. Estimations on the model and the immediate reward are then used to compute the $Q(s,a)$ values by iteratively solving the Bellman equation. The Hybrid Learner (Gläscher et al. [2010])is a weighted average of SARSA(0) and the Forward Learner. The Monte Carlo algorithm [Sutton and Barto, 2018] estimates the $Q(s,a)$ values in a model-free and incremental way like SARSA, but only at the end of the episode and using the within episode's return R_t, i.e the accumulated discounted reward from each state onwards. In the actor-critic algorithm, the V values are estimated by the critic and a RPE is fed into the actor to modify the policy parameters. The REINFORCE algorithm [Williams, 1992, Sutton and Barto, 2018] estimates the policy parameters of all the preceding within-episode decisions directly with gradient ascent using the return, also in a model-free manner.

A.2.3 Data Analysis

We fit algorithms to behavior using the Metropolis-Hasting Markov chain Monte Carlo (MCMC) method. Additionally, in order to avoid overfitting to behavior we perform 7-fold cross-validation. At each fold we leave 3 subjects out of the fitting procedure and we estimate the algorithms' parameters on the data of the remaining subjects. With the obtained parameter values we then assess the goodness of fit of the left-out subjects.

We acquired functional data of 23 subjects (11 female) on a 3T Siemens Prisma MRI Scanner, using a T2*-weighted 2D echo planar imaging (EPI) sequence. Two subjects were excluded from the analysis, one due to high degree of movement artifacts in his brain images and one due to performance (completion of less than half of the number of episodes that subjects performed on average). We performed preprocessing and statistical analysis of the fMRI images using the SPM12, SnPM13 and Nilearn software.

A.3 Results

A.3.1 De-correlation of RPE and SPE

After the experiment we fitted the SARSA(λ) and the Forward Learner algorithms to subjects' behaviour, obtained their corresponding RPE and SPE values and validated our experimental design. Figure A.3[a] depicts the values of the RPE and the SPE for one representative subject. Values that correspond to surprise trials are marked in red. The Pearson correlation coefficient of RPE and SPE without surprise trials (blue values) is $r = +0.453$. Adding controlled surprise trials to the experiment yields $r = +0.058$, namely an effective de-correlation. Figures A.3[b] and A.3[c] show the histograms of the RPE and the SPE respectively, at surprise trials for all subjects. The distribution of the SPE is shifted towards higher values, whereas the one of RPE is centered around 0.

The mean absolute RPE and SPE correlation across subjects is 0.09 ± 0.048 (mean ± std, 21 subjects). The maximum correlation observed was 0.15 and the minimum -0.143. Our experimental design successfully breaks the correlated gradual decrease of the two prediction errors over the course of learning.

A.3.2 Behavior: Actions close to goal are learned faster

The goal of this analysis of the behavioral data, was to quantify the progress (measured as percent-correct) of performance at *each state*.

According to reinforcement learning theory, decisions at states close to the goal are learned faster than decisions further away from the goal. This is obvious for an algorithm like SARSA-0, where in the first episode only the single state-action pair immediately leading to the goal state is reinforced. But also if eligibility traces are used to propagate reward information back to earlier decision, there is still a decay, be it in form of the memory decay $lambda$ or the distance dependent discount $gamma$. Analyzing the behavioral data, we observed that this delay in learning at distal states is not only a property of the algorithms used to model the data, but we extracted it directly from behavior, without fitting a particular model.

As we have two start states and different paths leading to the goal, not every state is visited in every episode. Furthermore, as the graph has loops, subjects can visit the same state several times within the same episode. Because of these reason and in order to make the learning curves comparable across different sates, the x-axis in Fig. A.2 represents the n^{th} episode in which a state actually has been visited. If a state was visited several times during the same episode, we only consider the subject's action during the first occurrence.

We found that participants achieved high levels of performance over the course of the experiment. Moreover, their speed of learning how to act at each state was related to the distance from the goal, but also to the difference in "correctness" between the two available actions. That is, the the performance at states one action before goal reached the (arbitrary) 80%

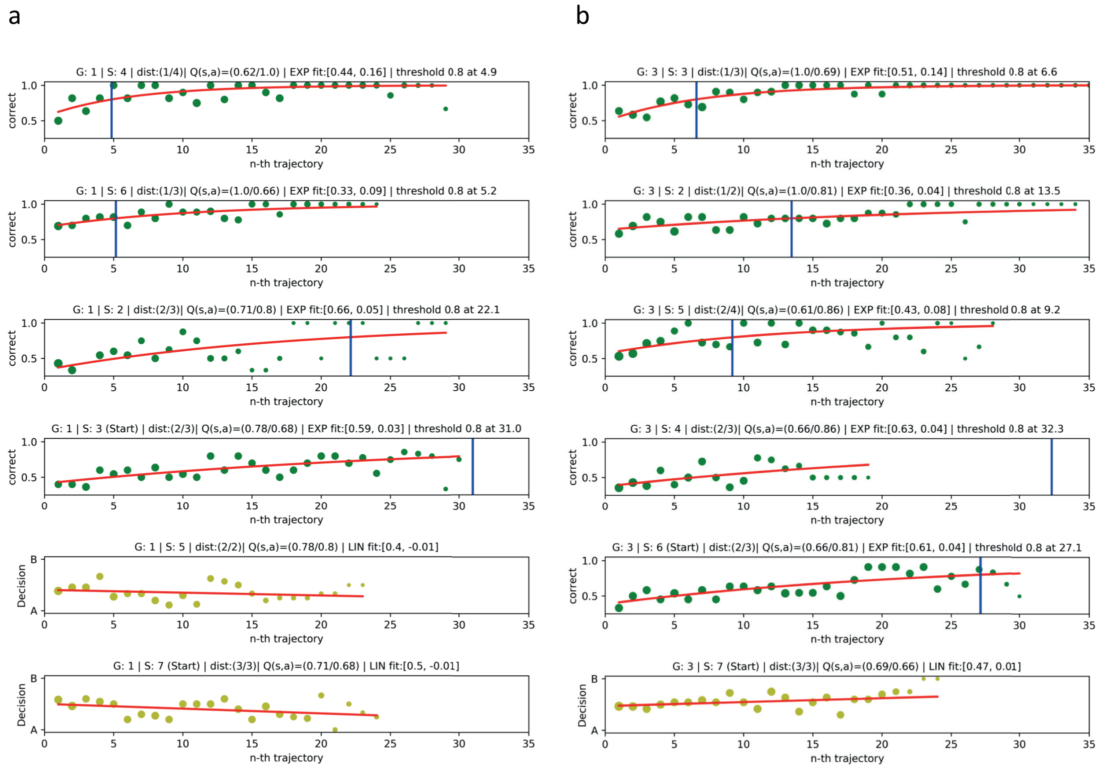


Fig. A.2. Learning curves per state. [**a**] Graph version G1. [**b**] Graph version G3. For both graphs, the learning curve at each state is shown in red. Green dots indicate the percent-correct score, averaged across all subjects. Dot size reflects the number of subjects (fewer subjects have done a large numbers of episodes). A saturating exponential curve (in red) was fitted to the dots (each dot weighted by number of subjects). The blue, vertical line highlights the trajectory count (n^{th} trajectory) when the red learning curve reaches the 80% percent-correct level. The figures in each column are sorted by *distance to goal* (states before goal are shown in the two top rows). The blue curves suggest that learning the correct action happens faster at states which are closer to the goal. The three states at the bottom (G1/S5, G1/S7 and G3/S7) do not have a "better" action. The distance to goal is the same for both. The yellow dots show the fraction of subjects that have chosen either action a or b. To the yellow dots a linear function was fitted. The information given in each figure title are: *G:* graph version. *S:* state ID. *dist(·/·):* distance to goal when selecting the correct action vs. distance to goal of the other action. *Q(s,a):* state-action values for the two actions. (correct action / alternative action). *EXP fit:* parameters u and v of the saturating exponential $1 - u \cdot exp(-TrajectoryCount/v)$. *threshold 0.8:* x-value of the blue line. *LIN fit:* parameters a and b of a linear fit $a + bx$.

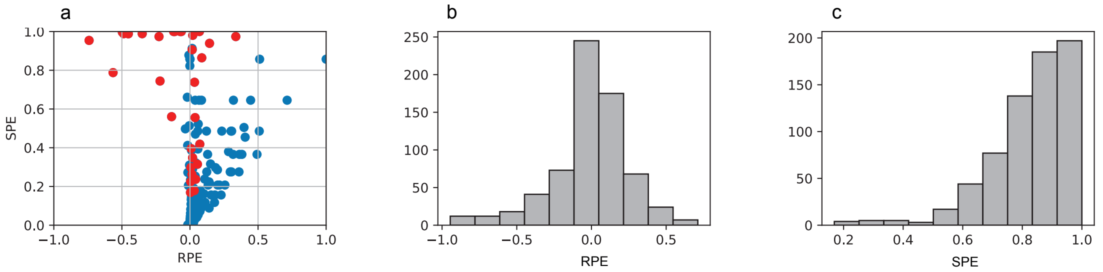


Fig. A.3. Post-hoc validation of the RPE/SPE de-correlation. **[a]** Each circle corresponds to a joint RPE/SPE value for each of the 192 actions of one representative subject. Surprise trials are indicated in red and normal trials in blue. Without surprise trials (blue values only), RPE and SPE are highly correlated (Pearson correlation coefficient r = +0.453). The addition of surprise trials to the experiment successfully de-correlated the two signals (r = +0.058). **[b], [c]** The two histograms show the distribution of RPE and SPE values respectively for all 675 surprise trials across all subjects. We observe that surprise trials have overall low RPE and high SPE values.

threshold in fewer trials than at states two actions before goal (Fig. A.2). Furthermore, there seems to be a trend for faster learning, if the difference between the two actions is larger. For example, in state S4 in graph G1 (Fig. A.2[a]) the correct action is learned very fast, as it leads to the goal state, while the wrong action leads to a state which is 4 steps away from goal (encoded as 1/4 in the Figure annotation).

A.3.3 Learning algorithms and behavioral fit

Table A.4 depicts the results of the model fitting procedure to subject's actions (post-experiment), in terms of Log Likelihood (LL) and Akaike Information Criterion (AIC). The purely MB Forward Learner seems to not be appropriate for this relatively large state-space task, and the Hybrid algorithm fits the data worse than purely MF algorithms. Next in the ranking come the MF Q-value estimators SARSA(λ) and Monte-Carlo. The actor-critic and the REINFORCE algorithm are the most likely models for behavior. In the actor-critic the resulting learning rate α_c of the critic was 0.00025, indicating that the critic had effectively no contribution in learning. It can thus be neglected, favouring an actor-only algorithm like REINFORCE. In this strategy, the policy parameters of the whole history of previous actions are updated upon receipt of immediate reward, modulated by a memory trace.

A.3.4 Neural signatures of multiple learning algorithms

After fitting the algorithms to the behavioral data we compute subjects' trial-by-trial learning signals and use them as regressors against the fMRI data in a general linear model. We find correlates of the MF RPE in the system that has been previously implicated in reward processing, including the ventral striatum, hippocampus, medial temporal lobe, ventromedial prefrontal cortex, and anterior and posterior cingulate cortex (A.5, top row). For the MB SPE we find correlated activity in the insula, inferior frontal gyrus, supplementary motor area and parietal lobe (A.5, middle row), in agreement with previously reported results (Gläscher et al. [2010]).

Algorithm		LL	#Param	AIC
Forward Learner	$\delta_{SPE} = 1 - T^a_{ss'}$ $T^a_{ss'} \leftarrow T^a_{ss'} + \alpha\, \delta_{SPE}, \quad s_{t+1} = s'$ $T^a_{ss''} \leftarrow T^a_{ss''} - \alpha\, T^a_{ss''}, \quad s'' \neq s'$ $Q_{MB}(s,a) = \sum_{s'} T^a_{ss'}(r + \gamma\ \max_{a'} Q_{MB}(s',a'))$	-1603.8	4	3215.6
Hybrid Learner	$Q(s,a) = w_t\, Q_{MB}(s,a) + (1-w_t)\, Q_{MF}(s,a)$ $w_t = w_0\, e^{-kt}$	-1390.3	7	2794.6
Monte Carlo	$Q(s,a) \leftarrow Q(s,a) + \alpha\,(R_t - Q(s,a))$	-1368.0	4	2744.0
SARSA(λ)	$\delta_{RPE} = r + \gamma\, Q_{MF}(s',a') - Q_{MF}(s,a)$ $Q_{MF}(s,a) \leftarrow Q_{MF}(s,a) + \alpha\, \delta_{RPE}\, e(s,a)$ $e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1 & , \ if \quad s = s_t, a = a_t \\ \gamma\lambda e_{t-1}(s,a) & , \ otherwise \end{cases}$	-1356.5	5	2723.0
Actor-critic	$\delta_{RPE} = r + \gamma\, V(s') - V(s)$ $V(s) \leftarrow V(s) + \alpha_c\, \delta_{RPE}\, e^c(s)$ $e^c_t(s) = \begin{cases} \gamma\lambda e^c_{t-1}(s) + 1 & , \ if \quad s = s_t \\ \gamma\lambda e^c_{t-1}(s) & , \ otherwise \end{cases}$ $p(s,a) \leftarrow p(s,a) + \alpha_a\, \delta_{RPE}\, e^a(s,a)$ $e^a_t(s,a) = \begin{cases} \gamma\lambda e^a_{t-1}(s,a) + 1 - \pi(s,a) & , \ if \quad s = s_t, a = a_t \\ \gamma\lambda e^a_{t-1}(s,a) & , \ otherwise \end{cases}$	-1330.1	7	2674.2
REINFORCE	$\theta \leftarrow \theta + \alpha\, \nabla \log \pi_\theta(s,a)\, R_t$	-1331.4	4	2670.8

Fig. A.4. Models and model selection Learning algorithms and their corresponding performance in explaining the behavioral data. Abbreviations and notations: LL: Log likelihood, #Param: number of parameters, AIC: Akaike Information Criterion, (s', a'): next state-action pair, r: immediate reward ($r = 1$ for the goal state, $r = 0$ for all other states), α: learning rate, γ: discount factor, $e(s,a)$, $e^c(s)$ and $e^a(s,a)$: exponentially decaying eligibility traces, $p(s,a)$: preference for action a when in state s, θ: policy parameter vector, ϕ: feature vector (1 for current (s, a), 0 otherwise), return $R_t = \sum_{k=1}^{T-t} \gamma^{k-1} r_{t+k}$. All algorithms were used in combination with a softmax action selection policy $\pi(s,a) = e^{f(s,a)/\tau} / \sum_b e^{f(s,b)/\tau}$, where $f(s,a) \in \{Q(s,a),\ p(s,a),\ \phi(s,a)^T\theta\}$ depending on the algorithm.

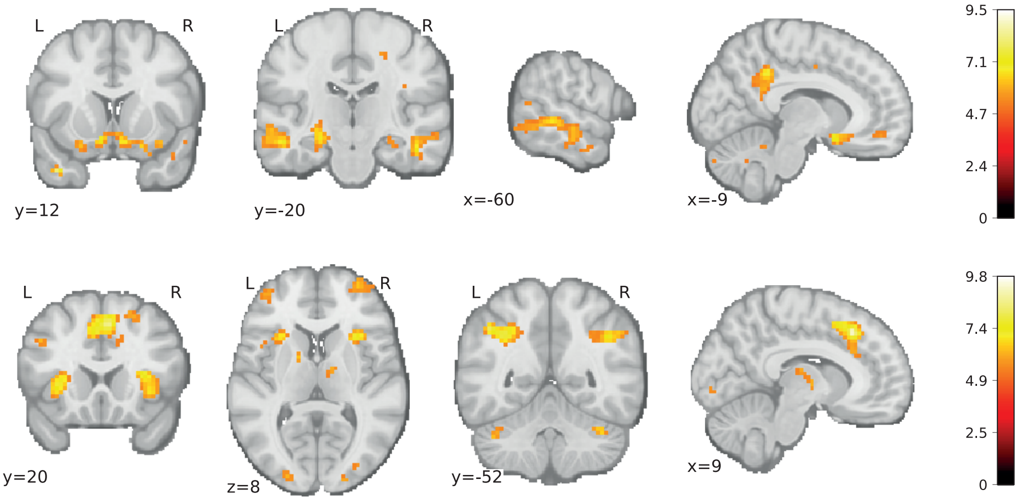


Fig. A.5. Neural correlates of RPE and SPE. T-statistic maps of: *Top Row*: RPE in ventral striatum, hippocampus, medial temporal lobe, ventromedial prefrontal and posterior cingulate cortex. *Bottom Row*: SPE in insula, inferior frontal gyrus, parietal lobe and supplementary motor area. [21 subjects, random effects analysis, whole brain family-wise error (FWE) correction p<0.05, nonparametric permutation test]

A.4 Discussion

We have designed and employed a multi-step sequential decision making task to dissociate and track human brain RL signals that so far have been mostly studied in two-stage tasks (but see [Simon and Daw, 2011] for a spatial navigation task, with no de-correlation). We found that earlier results generalized to a multi-step setting. In particular, our results support the existence of two systems in the brain involved in MF RPE-mediated value learning and MB computations, in agreement with previous studies (Gläscher et al. [2010]). Additionally, our behavioral data hint at a possible third system acting only on the policy and involving the reinforcement of sequences of actions preceding reward. Although the MB Forward Learner did not explain behavior well, we found correlates of an SPE. This suggests that an MB learning system was active, possibly building an internal model of the task, but it was not (yet) in control.

In ongoing work, we investigate the neural representations of three state classes (start, goal, before goal). Preliminary results using multivariate pattern analysis (MVPA) indicate that the states can be classified from the fMRI recordings. Furthermore, analyzing the high dimensional voxel data shows a clustering of the class representations. These representations are more stable for start and goal classes than for the before-goal class.

B Adaptive control between multiple reinforcement learning systems

Here, we study human learning in a complex environment, where only a single action at each state brings the participant closer to the goal, while all other actions lead back to *trap states*, far from the goal. Classic reinforcement learning models (model-free and model-based) fail to learn the task. That is, the absence of external reward leads to very long exploration in the first episode. On the other hand, we found that humans could recognize traps and avoid them quickly. Inspired by human behavior, we developed a hybrid model which combines both model-free and memory-based learners, mediated by an internal feedback signal (the "novelty of states"), and managed to achieve human performance level. We also tested the new hybrid model on knowledge-transfer tasks, we showed that the memory-based learner can serve as a prior in a new, but similar environment, which facilitated learning. EEG N1 amplitudes were significantly increased in the new environment compared to the original environment, showing the evidence of the knowledge updating process, which is in line with the model prediction.

Contributions

Designed the experiment: He Xu, Michael Herzog
Implemented the experiment: He Xu
Ran the EEG experiments: He Xu, Marco Lehmann
Analyzed the behavioral data: Marco Lehmann, He Xu
Analyzed the EEG data: He Xu
Discussed and interpreted the results: He Xu, Marco Lehmann, Michael Herzog
Discussed and developed the novelty-based learning model: He Xu, Marco Lehmann
Wrote the manuscript: He Xu, Michael Herzog

This is work in progress, led by He Xu. The following manuscript summarizes preliminary results.

B 1 Introduction

Humans can learn from delayed and sparse feedback. For example, in chess a reward is issued (win, loss, tie) only after many moves, which makes it impossible to evaluate the goodness of a single move. Reinforcement learning (RL) theories offer a variety of explanations how learning can occur in these situations. The models operate often in an environment, comprised of states and actions, through which an agent or human navigates. At each state an action is made, which brings the human or agent to the next state until a goal state is reached (Figure 1).

There are two main types of RL models, namely, model-free and model-based models. The basic idea of model-free learning is that to each state-action pair a value is assigned, which tells how good the state-action pair is. This value is updated whenever a state-action pair is executed. In general, when the actual reward is larger than expected, a positive reward prediction error (RPE) is produced and the value is increased. When the actual reward is lower than expected, a negative RPE is produced and the value is decreased. Hence, the RPE drives the learning process. Using the values of states or state-action pairs, the model-free learner can learn to find its way to the goal.

Different from model-free learning, model-based learners build up a model of the environment by learning the state-action transitions. The state-action transition records the probability of going to another state when taking an action at the current state. Whenever a state is visited, the current state-action transition is updated, based on the state prediction error (SPE), which measures the difference between the current state-action transition model and the observed transitions. Using these transitions, the model-based learner can plan its way to the goal.

Behavioural experiments have shown that humans use both model-free and model-based learning. fMRI and EEG studies showed evidence that humans use the RPE (Bellebaum & Daum 2008; Gehring & Willoughby 2002; Gruendler et al. 2011; Tucker et al. 2003; Badgaiyan & Posner 1998; Cohen et al. 2007; Nieuwenhuis et al. 2005; Doñamayor et al. 2011; Haruno & Kawato 2006; McClure et al. 2003; O'Doherty et al. 2003) and the SPE (Courchesne et al. 1975; FABIANI & FRIEDMAN 1995; Opitz et al. 1999; Glaescher et al. 2010). It is an unknown question when humans use model-free and model-based learning and how they are related. It was shown that time pressure, cognitive demands, instructions and task specific costs/benefits, can change the balance between the two RL learning models (Doll et al. 2009; Kool et al. 2018; Kool et al. 2017; Otto et al. 2013; Keramati et al. 2016). For example, humans showed more model-based behaviour whenever cognitive resources are available, and if under cognitive load, they tend to behave in a model-free manner.

The main question is how model-free and model based learning can be integrated into one framework and to determine under which condition either one prevails. To the best of knowledge, there is only one such hybrid model (Glaescher et al. 2010). In this model, the trade-off between the two systems was determined by a free, meaningless parameter (Camerer & Ho 1998).

Here, we propose that novelty is a feedback itself, which leads to model-based learning and determines whether model-based or model-free learning prevails. We present a hybrid model, which uses novelty and show evidence for the model by a behavioural and imaging (EEG) experiment. In particular, we show that swapping two states in the environment leads to updating in the model-based learner and is well reflected in the N1 component of the EEG.

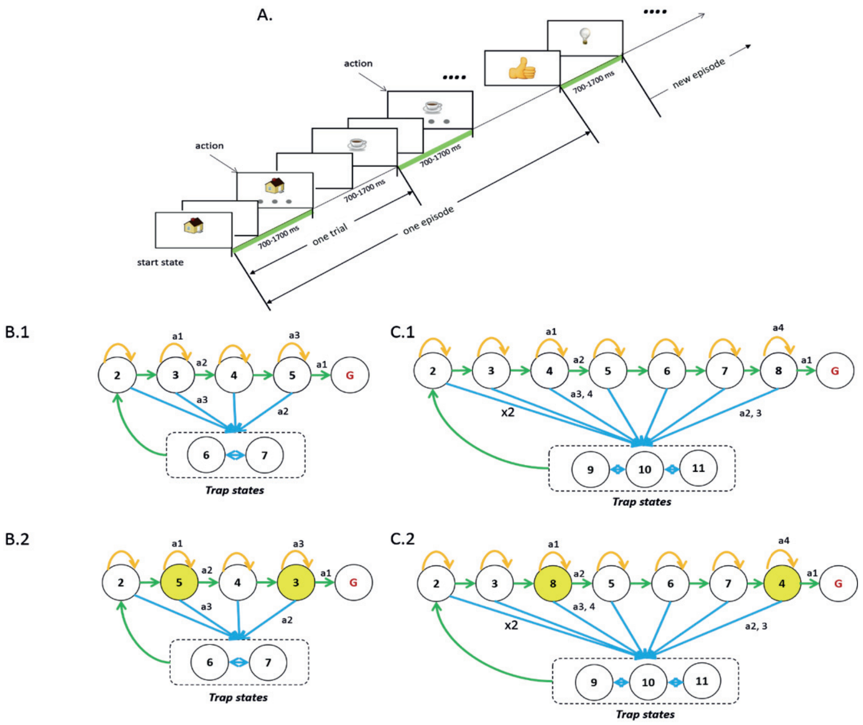


Figure 1 (A) After an image (state) was presented, participants needed to wait for 700-1700ms, randomly chosen, until grey disks (actions) were presented at the bottom of the image. After a click (action), a blank screen was shown for 700 and 1700ms, randomly chosen, and then the next image appeared. 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. The structures of the simple and complex environment. Non-goal states are presented using numbers; the goal state is presented as the red G; trap states are highlighted by dashed rectangles. Actions are presented by arrows. Green arrows present the actions that bring participants closer to the goal state. Yellow arrows present actions that let participants stay at the current state. Blue arrows present the actions that bring participants to the trap states. For example, there are 7 states in the simple environment, including (1) progressing states (images 2, 3, 4, 5), (2) the goal state (red G), and (3) the trap states (images 6, 7). Participants can make 3 possible actions at each state: one action (green arrow) brought them to the next progressing state, one action (blue arrow) brought them to one of the trap state, one action (yellow arrow) let them stay at the current state. After 5 episodes, image 3 and image 5 are swapped while the actions remained the same as before.

B 2 Results

B 2.1 A hybrid model based on novelty

Here, we developed a framework that combines model-based and model-free RL learners (Figure 2). We refer to the model-based RL learner as a "Memory-based learner" and the model-free RL learner as a "Value-based learner". We propose that novelty is the balancer between the two. If a participant visits a state many times, novelty decreases; if a participant visits a state rarely, novelty is high. We define the novelty as:

$$Novelty\ (\boldsymbol{state}) := \frac{1}{number\ of\ visits\ to\ the\ \boldsymbol{state}}$$

The value-based learner is implemented using SARSA(λ) (Sutton et al. 1998), which learns the state-action values based on external reward (see Appendix). The reward prediction error (RPE) is computed to update state-action values. The value-based learner produces action selection policies to maximize the total amount of external reward it can obtain. The memory-based learner learns not only the transitions between states, but also the internal outcomes of states. The state transitions are updated based on observations. The internal outcome is modulated by the novelty of the state, and also by the value of the state (Figure 2). If the value-based learner does not assign a value to a state, the outcome is alongside with the novelty of the state. If the state value is positive, the outcome of the state will count for both novelty and value of the state. However, if the state value is negative, the outcome of the state is adverse to the novelty of the state. The policy produced by the memory-based learner aims to maximize the outcomes it can obtain.

The policies of the value-based and memory-based learner may differ strongly and hence a combined policy is needed. In our model, this policy is computed as:

$$Policy_{hybrid}(S) = novelty(S) \times Policy_{mbl} + [1 - novelty(S)] \times Policy_{vbl}$$

**mbl = memory-based learner, vbl = value-based learner*

With this model, we predict that the learning could be accomplished, meaning the optimal actions will be found, in the first episode even when no reward is found. With low novelty in the trap states (since they will be frequently visited), learning will be slow and little. Learning will be faster in the progressing states when getting closer to the goal. We also predict that if a simple environment is perturbed, re-learn will occur. And if a complex environment is perturbed, only updating is needed. The difference between re-learn and updating process can be shown from model simulation and electrophysiology signals. Since the N1 component in EEG is considered to be reflecting the belief updating process (Friston et al. 2017), we expect to see that the N1 component changes for the updating process (in perturbed complex environment) but not for the re-learn process (in perturbed simple environment).

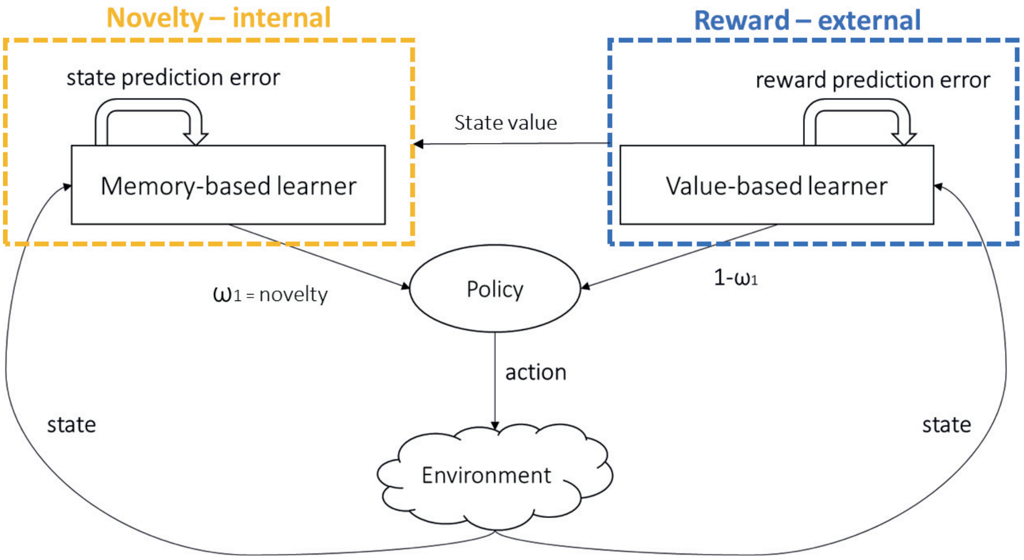


Figure 2 The model consists of two parts: the value-based learner and the memory-based learner. The value-based learner learns the reward function given by the environment in a model-free manner. The memory-based learner learns the state transitions and the novelty of states from internal experience. The value-based learner also passes the state value function to the memory-based learner. Both learners propose their policies to the policy controller. The policy controller evaluates the policies by the novelty of the states. Detailed implementation is presented in box 1-4.

B 2.2 A combined behavioural and EEG Experiment

We tested to what extent participants use model-based and model-free learning in a sequential decision make paradigm developed previously (Figure 1; Tartaglia et al. 2017).The learning environment contained three types of states: progressing states, trap states, and a goal state. Progressing states are the "good" states that bring participants to the goal. Trap states are the "bad" states that led participants away from the direct path to the goal. At each state, there were three types of actions that participants could choose: (1) actions that bring participants to a progressing state (green arrows in Figure 1), (2) actions that let participants stay at current state (yellow arrows in Figure 1) and (3) actions that bring participants to a trap state (blue arrows in Figure 1).

We used two environments (Figure 1). Participants explored the simpler environment with 7 images for 5 episodes (Figure 1). Next, we swapped two images (Figure 1) without notifying the participants. Participants explored the modified environment for another 5 episodes. After a short break of 3-5min, participants explored the more complex environment with 11 states (Figure 1). After 5 episodes, two states were swapped and participants continued for another 5 episodes. New images were used for the complex environment. The starting images of the 5 episodes in the swapped environment were the same as before swapping. We used different images for different participants. EEG was recorded during the entire experiment.

Behaviour Results

It took participants on average 45 actions to find the goal state in the first episode of the simple environment and 117 actions in the complex environment. Performance strongly improved from episode 2 on for both environments quickly reaching optimal performance (Figure 3). In particularly, observers learned to avoid the trap states. We analysed performance in the first episode in detail (Figure 3). In the simple environment, optimal actions were found for all states, i.e., even before the goal was found for the 1st time. In the complex environment, optimal actions were found for the progressing states but not for the trap states. The closer the progressing state is to the goal, the faster the optimal action was found.

After episode 5, we swapped 2 images. In the simple environment, performance deteriorated to a performance level, which is close to the one in the first episode. In the complex environment, there was a significant improvement between the performance level in the 1st episode and the 1st episode after swapping. Performance improved again in both environments reaching optimal performance.

First, model-free learners cannot easily explain our results because there is no learning in episode 1, i.e., before the goal state was found, contrary to human learning (Figure 3C). Second, model-based learners are also challenged by our data because trap states are not recognized by the model but by the participants (Figure 3). Third, the data can be well explained by our hybrid model because in trap states the novelty is low, hence, the weight on the memory-based learner is low, and participants rely on the value-based learner. For this reason, actions are chosen randomly. In progressing states, the novelty is high, hence, participants rely more on the memory-based learner and actions are chosen to avoid going to low-novelty states (trap states). Fourth, both the model-based and the model-free learners were seriously deteriorated when we swapped the images after episode 5 and stayed deteriorated for the following episodes. Performance is deteriorated because the previously learned state-action transitions and the Q-values are inappropriate. The hybrid-model does a good job in this structure because it can still identify the trap states and avoid it. The model can, thus, restrict re-learning to fewer states.

EEG results

We determined the Event-Related Potentials (ERPs) for the two swap states focusing on the N1 component, which is known to reflect updating processes (Friston et al. 2017). In the simple environment, there were no obvious changes after swapping the states (t(10) = 0.93, p = 0.37). However, in the complex environment, we found clear increases of the N1 for all swapped and non-swapped states (t(18) = 2.48, p =0.02), indicating a global updating process. We suggest that in the simple environment participants relearn the entire structure, whereas in the complex environment they only update the swapped states (Figure 1C). As a control, we analysed the N1 amplitudes in the first two and last two episodes before swapping. There were no differences in the N1 (simple environment: t(10) = -1.00, p = 0.33; complex environment: t(18) = 0.90, p = 0.37), indicating that the N1 is sensitive to updating but not learning in general.

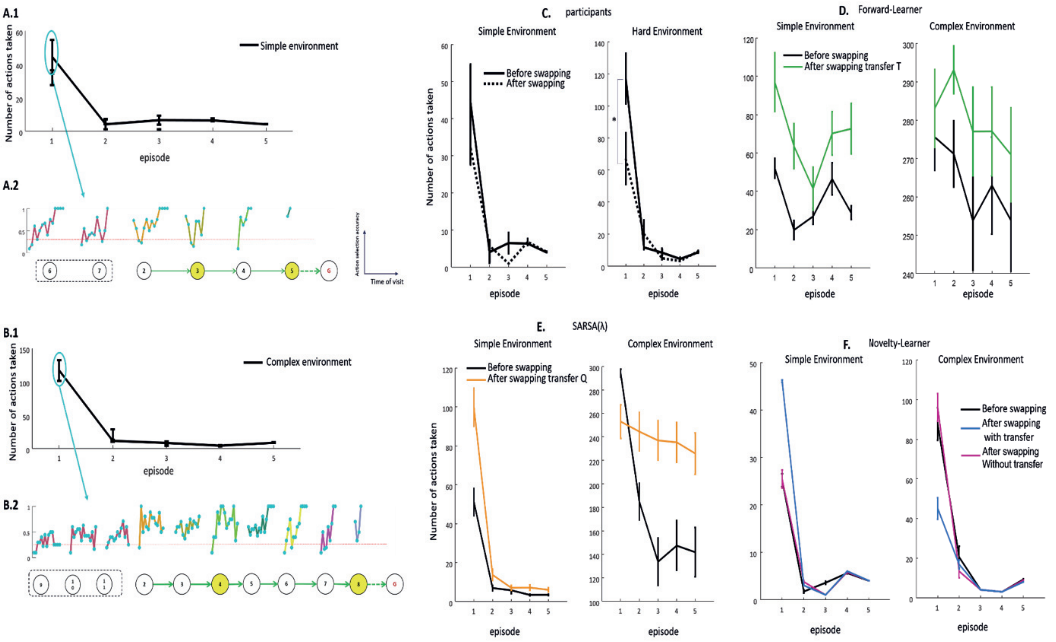


Figure 3 (A.1, B.1) Learning curves for the first 5 episodes of the simple and complex environment before swapping images. The x-axis presents the number of episodes completed. The y-axis presents the number of actions taken in each episode before the goal was found. Error bars show the standard error of the mean (SEM). **(A.2, B.2)** Learning curves in the first episode before swapping the images. Each blue point on the learning curve represents one visit to a state. The chance level of choosing the correct action (green arrows in Figure 1) is 1/(number of actions per state), which is 1/3 for the simple environment and 1/4 for the complex environment, presented by red dashed lines. The action selection accuracy of each visit is averaged over all participants, presenting how likely the participants choose the optimal action. Growing action selecting accuracy means that participants learn the optimal actions over visits. Fewer visits to a state (fewer blue points) means participants learned the correct action faster. **(C-F)** Learning curves of participants and models for the two environments before and after swapping the images. The x-axis presents the number of episodes. After swapping the images, the episode was counted start from 1 again. The y-axis presents the number of actions taken to finish one episode. **(C)** Participants were able to find the optimal actions in the first episode and performance were largely improved from episode 2 on. The performance was not significantly improved after swapping the images in the simple environment, but was significantly improved in the complex environment. **(D)** Forward-Learner agent performance was worse than human participants in both environments. **(E)** SARSA(λ) agent performance was worse than human participants in both environments. **(F)** Novelty-learner agent was able to achieve human performance level.

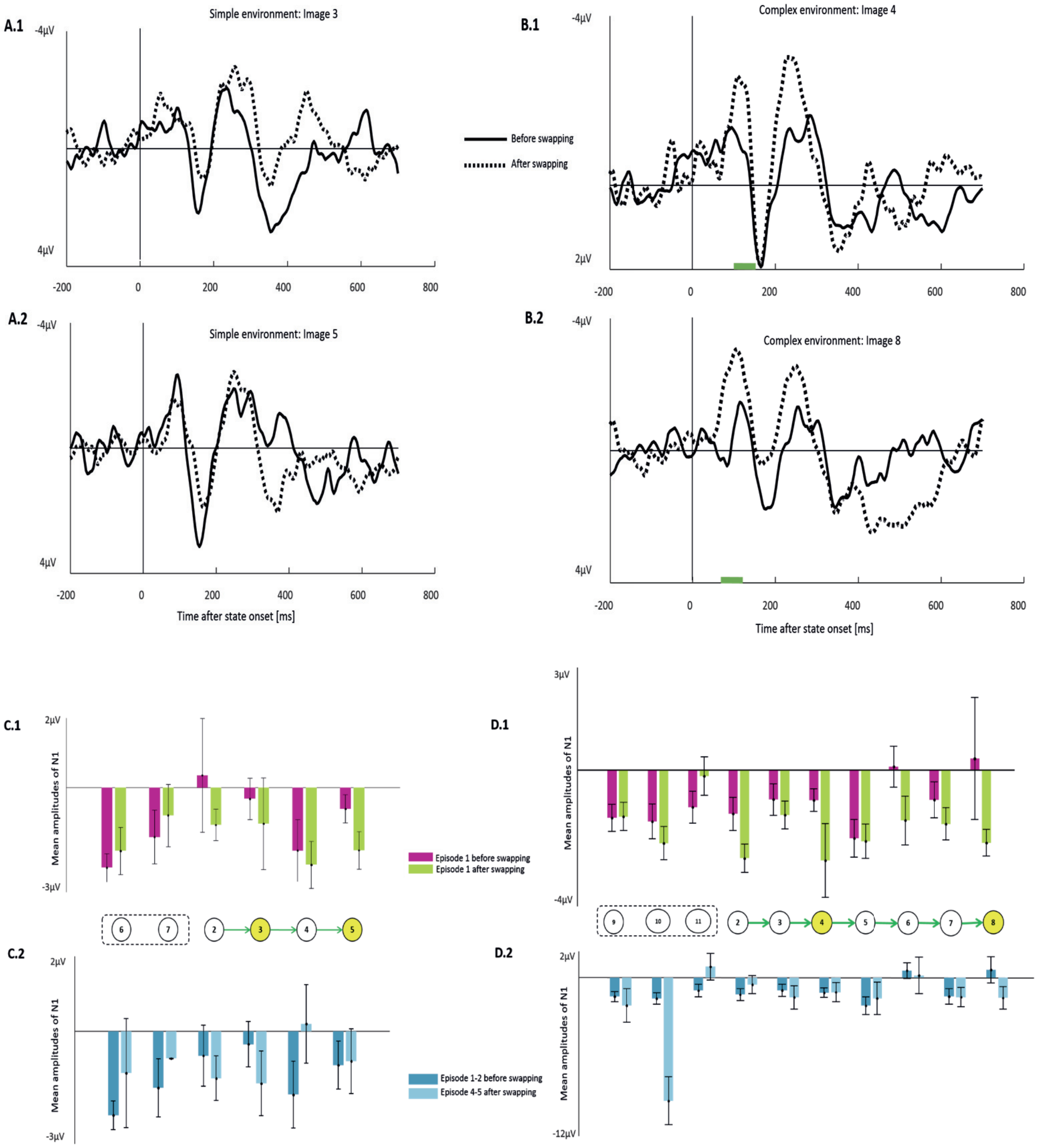


Figure 4 (A-B) The ERPs of the swapped images before and after swapping. 0 on each x-axis indicates time of image onset. Negative values are plotted up by convention. Green windows indicate significance.

(A1-2) The swapped images were image-3 and image-5 in the simple environment. The ERPs of the two images did not show significant difference between before and after swapping (Image 3: t(20) = 0.47, p = 0.63. Image 5: t(20) = -0.18, p = 0.85).

(B1-2) The swapped images were image-4 and image-8 in the complex environment. The N1 mean amplitudes were significantly larger after than before swapping (Image 4: t(20) = 2.6, p = 0.01. Image 8: t(20) = 2.11, p = 0.04).

(C-D) The comparison of N1 mean amplitudes in the learning process and the updating process.

(C-1) The simple environment. The N1 amplitudes of each state in the first episode before and after swapping. State 6,7,2,4 and the goal were presented by the same images in all episodes. The images presenting state 3 and 5 were swapped after 5 episodes. The N1 mean amplitudes were not significantly different before and after swapping (t(10) = 0.93, p = 0.37).

(C-2) The simple environment. The N1 amplitudes of each state before swapping, compared between the early episodes (episode 1-2) and the late episodes (episode 4-5). The amplitudes difference was not significant (t(10) = -1.00, p = 0.33).

(D-1) The complex environment. The N1 amplitudes of each state were significantly different before and after swapping. All states except state 4, 8 were presented by the same images in all episodes. The N1 amplitudes were significantly different for all states (t(18) = 2.48, p = 0.02).

(D-2) Experiment 2. The N1 amplitudes of each state before swapping compared between the early episodes (episode 1-2) and the late episodes (episode 4-5). The amplitude difference was not significant (t(18) = 0.90, p = 0.37).

B 3 Methods

B 3.1 Experiment set up

Experiments were conducted on a Phillips 201B4 monitor, running at a screen resolution of 1,980 × 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 (Brainard 1997).

B 3.2 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 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.

B 3.3 Stimuli and general procedure

Before starting the experiment, we showed the participants the goal image that they were required need to find. Next, participants were presented the other images that they may encounter during the experiment. After familiarizing with the images, participants clicked the 'start' button to start the experiment proper. At each trial, participants were presented an image (state) and a number of grey disks below the image (Figure 6). 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. The next episode started with an image chosen randomly (except for that the goal image was never the first image).

B 3.4 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 (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, a time window (green interval in Figure 1A) was extracted from 200ms before to 700ms after the image onset. The baseline activity was removed 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.

B3.5 Model implementation

For our model simulation, we use the standard *SARSA(λ) algorithm for the* model-free model and the forward learner for the model-based models. In principle, other models could be used too. Even though both models are standard, we quickly describe the implementations we used.

B 3.5.1 Model-free agent: SARSA(λ)

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

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

where γ is a discounting rate parameter to determine the present value of future rewards. 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 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(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:

Equation 3

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

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

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.

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

B 3.5.2 Model-based agent: Forward-Learner

The model-based RL agent we chose to use is the Forward-Learner model (Gl??scher et al. 2010) which estimates the transition probability from state s to next state s′ via action a, defined as T(s, a, s′). After observing the next state s′, the Forward-Learner agent computes a state prediction error (SPE) by:

$$SPE_t = 1 - T(s, a, s')$$

Then the agent updates the transition probability by:

$T(s,a,s') = T(s,a,s') + \eta \times SPE_t$ for visited s′

$T(s,a,s'') = T(s,a,s'') \times (1-\eta)$ for un-visited s′′

where η presents the learning rate of the agent. To learn the reward function r(s) in the environment, the Forward-Learner agent also computes the state-action values Q(s,a) by:

$$Q(s,a) = \sum_{s'} T(s,a,s') \times (r(s') + maxQ(s',a'))$$

The Q-values are used to generate action selection policy in the same way shown in Eq.4.

B 3.5.3 Novelty leaner

In our model framework, we assume that human participants used both model-based and model-free RL systems. From the behavior results, we also assumed that humans controlled the weights between the two RL systems adaptively through an interval signal, which we proposed to be the novelty of states.

The novelty-learner agent contains two learners: a value-based learner and a memory-based learner. The value-based learner learns the external reward given by the environment, and can be implemented as the classical model-free models such as SARSA(λ).

The memory-based learner is a model-based agent, which learns the transitions between states. It also computes the novelty of a state based on the frequency of visits as the following equation:

$$Novelty\ (\boldsymbol{S}) = \frac{1}{number\ of\ visits\ to\ \boldsymbol{S}}$$

The novelty of a state serves as an interval feedback signal that takes part in the policy evaluation. Based on the novelty of a state, the memory-based learner also estimates the state's interval value, which we define as the outcome of the state. The outcome of a state does not only depend on its novelty, but also depend on the state value estimated from the value-based learner. With this implementation, memory-based learner is able to distinguish between "new and good" states (states 2-8 in Figure7) and "new but bad" states (states 12-13 in Figure 7). The outcome of a state is defined as:

$$Outcomes(\mathbf{S}) = Novelty(S) \times Outcomes(\mathbf{S}) + [1- Novelty(S)] \times \mathrm{V}(\mathbf{S})$$

After visiting a state, both the value-based learner and the memory-based learner propose their action selection policy to the policy controller. The policy controller computes the final policy based on the novelty of the state. If a state has a high novelty, the final policy will be biased towards the memory-based learner's policy. Otherwise, it will be biased towards the value-based learner's policy. The final policy is computed by:

$$Policy_{hybrid}(S) = novelty(S) \times Policy_{mbl} + [1 - novelty(S)] \times Policy_{vbl}$$

**mbl = memory-based learner, vbl = value-based learner*

The detailed implementations of the memory-based learner and policy controller are presented in box 1-4.

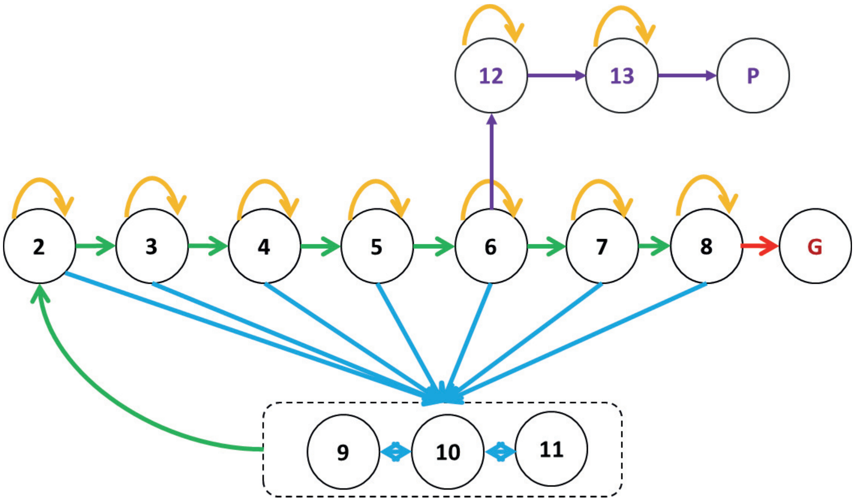
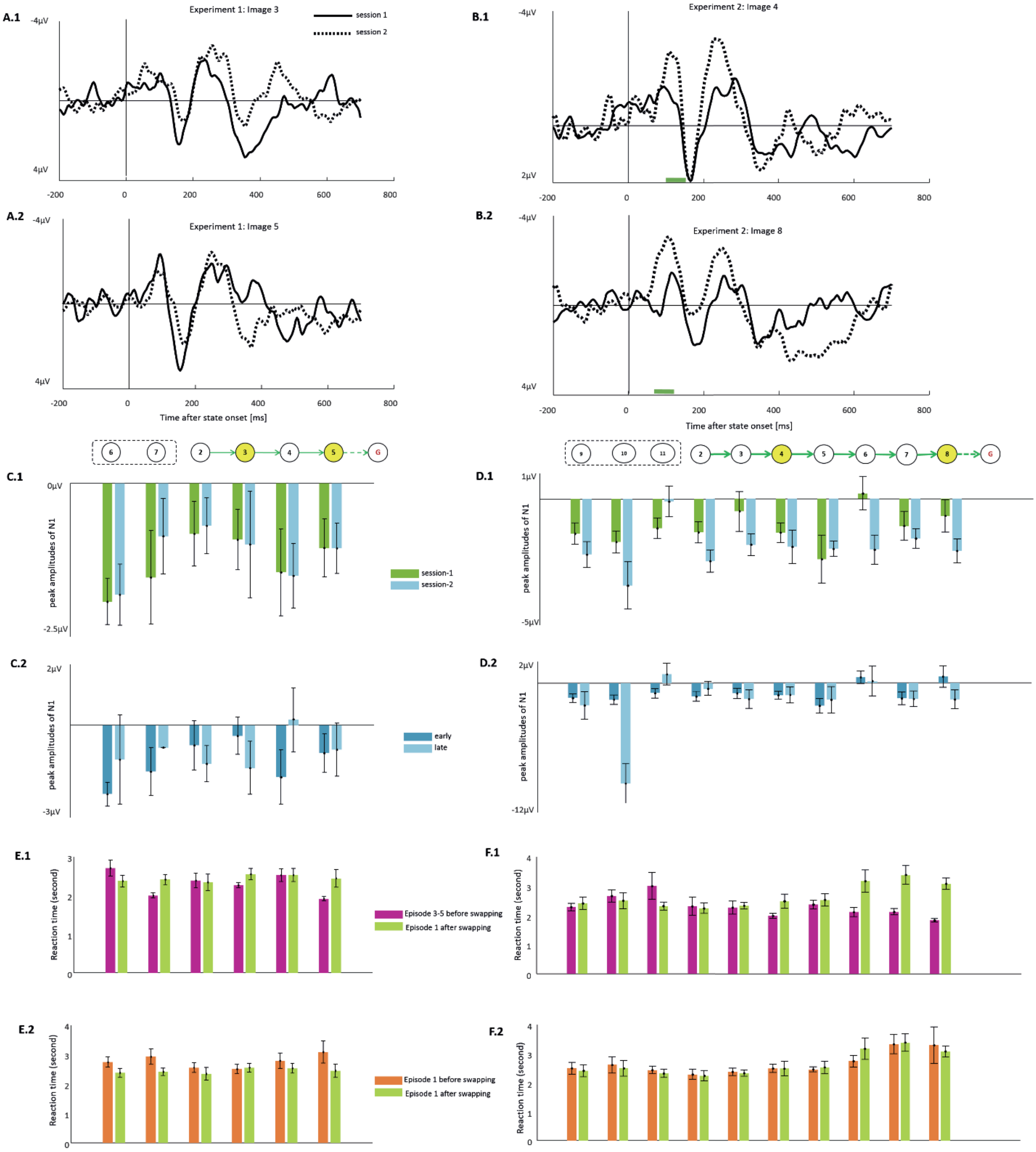


Figure 5 An alternative design of the environment, which contains both a reward and a punishment. The reward state is presented by the red G. The punishment state is presented by the purple P. In order to collect more reward in this environment, participants should avoid going to the purple states and the trap states.


A.1
Experiment 1: Image 3
session 1
session 2
-4µV
4µV
-200
0
200
400
600
800
A.2
Experiment 1: Image 5
-4µV
4µV
-200
0
200
400
600
800
Time after state onset [ms]
B.1
Experiment 2: Image 4
-4µV
2µV
-200
0
200
400
600
800
B.2
Experiment 2: Image 8
-4µV
4µV
-200
0
200
400
600
800
Time after state onset [ms]
6
7
2
3
4
5
G
C.1
0µV
-2.5µV
peak amplitudes of N1
session-1
session-2
C.2
2µV
-3µV
peak amplitudes of N1
early
late
9
10
11
2
3
4
5
6
7
8
G
D.1
1µV
-5µV
peak amplitudes of N1
D.2
2µV
-12µV
peak amplitudes of N1
E.1
Reaction time (second)
0
1
2
3
Episode 3-5 before swapping
Episode 1 after swapping
E.2
Reaction time (second)
0
1
2
3
4
Episode 1 before swapping
Episode 1 after swapping
F.1
Reaction time (second)
0
1
2
3
4
F.2
Reaction time (second)
0
1
2
3
4

B 3.6 Model Simulation procedure

We ran the simulation of each model over their parameter space in order to estimate the model's best performance.

For the Forward-Learner model, its parameter space contains two parameters – η (learning rate) and τ (temperature). For each parameters, we took 10 values from 0.01 to 1.0 (0.01, 0.11, 0.21, 0.31, 0.41, 0.51, 0.61, 0.71, 0.81, 0.91, 1.0). We tried all the 100 combinations of the two parameters for the model simulation.

For the SARSA(λ) model, its parameter space contains four parameters – α (learning rate), γ (discounting rate), λ (eligibility trace), τ (temperature). We took 10 values form 0.01 to 1.0 for each parameter and ran the model simulation for each parameter combination. The model performance was evaluated in the same way as the Forward-Learner model. For the novelty-learner model, its parameter space contains six parameters – four parameters from the value-based learner (SARSA(λ) model) and two parameters from the memory-based learner.

For each parameter set in each model, we ran the simulation for 50 times and estimate the average number for actions needed to finish 5 episodes. We selected the parameter set which finished the 5 episodes with the lowest number of actions. This parameter set was used to generate the model's best performance shown in Figure 4.

Implementation of Memory-based Learner

Memory-based learner initialization (for deterministic environments)	
Input:	nS – number of states nA – number of actions D – decay rate (free parameter) T – temperature (free parameter)
Implementation:	1. Initialise the ***transition_matrix*** of size $nS \times nA$ with all zeros 2. Initialise the states' ***visit_counter*** of size $1 \times nS$ with all zeros 3. Initialise the states' ***outcomes*** of size $1 \times nS$ with all ones 4. Set the outcome of the goal state bigger than one (for example, 10) 5. Initialise the parameters as given in the input

Memory-based learner update (for deterministic environments)	
Input:	S: visited state A: action taken at state S S′: next state after taking action A at state S V(S′): value of state S′, given by the value-based learner

Implementation:

```
Case 1: if transition_matrix(S,A) = 0
        transition_matrix(S,A) = S′
        visit_counter(S′) = visit_counter(S′) + 1
        novelty(S′) = 1/ visit_counter(S′)
        outcomes(S′) = novelty(S′) ×outcomes(S′) + [1- novelty(S′)]× V(S′)

Case 2: if transition_matrix(S,A) = S′
        visit_counter(S′) = visit_counter(S′) + 1
        novelty(S′) = 1/ visit_counter(S′)
        outcomes(S′) = novelty(S′) ×outcomes(S′) + [1- novelty(S′)]× V(S′)

Case 3: if transition_matrix(S,A) ≠ 0 and transition_matrix(S,A) ≠ S′
        Reset all transitions from state S to zeros
        Decay the outcomes of all states by (1-D)
        transition_matrix(S,A) = S′
        visit_counter(S′) = 1
        novelty(S′) = 1/ visit_counter(S′)
        outcomes(S′) = novelty(S′) ×outcomes(S′) + [1- novelty(S′)]× V(S′)
```

Memory-based learner action selection (for deterministic environments)	
Input:	S: visited state T: temperature (free parameter)

Implementation:	Case 1: if for any action a at state S: ***transition_matrix***(S,a) = 0 action_preference(a) = 0, for all a Case 2: if for some action a at state S: ***transition_matrix***(S,a) = 0 action_preference(a) = 1, for a that ***transition_matrix***(S,a) = 0 action_preference(a') = 0, for a' that ***transition_matrix***(S,a) ≠ 0 Case 3: if for any action a at state S: ***transition_matrix***(S,a) ≠ 0 action_preference(a) = outcome(S'), for all actions a, and S' = ***transition_matrix***(S,a) $\boldsymbol{Policy\_MBL}(S,a) = \frac{\exp\left(\text{action\_preference}(S,a)/T\right)}{\sum_i \exp\left(\text{action\_preference}(S,i)/T\right)}$
Output:	***Policy_MBL*** (policy proposed by Model-based Learner)

Implementation of Policy Controller

Policy Evaluation	
Input:	Policy_MBL(S): policy proposed by Model-based Learner at state S Policy_VBL(S): policy proposed by Value-based Learner at state S novelty(S): novelty of state S
Implementation:	Policy_final = novelty(S) × Policy_MBL(S) + [1-novelty(S)] × Policy_VBL(S)
Output:	***Policy_final***

Model simulation parameters – Experiment 1

SARSA(λ)	
Session 1	$\alpha = 0.40$, $\gamma = 0.80$, $\lambda = 0.60$, $\tau = 0.21$
Session 2 – with Q transfer	$\alpha = 0.85$, $\gamma = 0.40$, $\lambda = 0.80$, $\tau = 0.81$
Forward-Learner	
Session 1	$\eta = 0.60$, $\tau = 0.21$
Session 2 – with Q transfer	$\eta = 0.85$, $\tau = 0.11$

Session 2 – with T transfer	$\eta = 0.75$, $\tau = 0.81$
Novelty-Learner	
Session 1	$\alpha = 0.35$, $\gamma = 0.60$, $\lambda = 0.90$, $\tau = 0.01$, $D = 0.15$, $T = 0.11$
Session 2 – with transfer	$\alpha = 0.05$, $\gamma = 0.5$, $\lambda = 0.5$, $\tau = 0.01$, $D = 0.05$, $T = 0.41$

Model simulation parameters – Experiment 2

SARSA(λ)	
Session 1	$\alpha = 0.80$, $\gamma = 1.00$, $\lambda = 0.70$, $\tau = 0.31$
Session 2 – with Q transfer	$\alpha = 0.90$, $\gamma = 0.30$, $\lambda = 0.50$, $\tau = 0.61$
Forward-Learner	
Session 1	$\eta = 0.95$, $\tau = 0.11$
Session 2 – with Q transfer	$\eta = 0.45$, $\tau = 0.81$
Session 2 – with T transfer	$\eta = 0.10$, $\tau = 0.01$
Novelty-Learner	
Session 1	$\alpha = 0.35$, $\gamma = 0.90$, $\lambda = 1.00$, $\tau = 0.21$, $D = 0.55$, $T = 0.91$
Session 2 – with transfer	$\alpha = 0.45$, $\gamma = 0.90$, $\lambda = 0.50$, $\tau = 0.01$, $D = 0.05$, $T = 0.71$

Badgaiyan, R.D. & Posner, M.I., 1998. Mapping the cingulate cortex in response selection and monitoring. *NeuroImage*, 7(3), pp.255–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9597666.

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

Camerer, C. & Ho, T.H., 1998. Experience-Weighted Attraction Learning in Coordination Games: Probability Rules, Heterogeneity, and Time-Variation. *Journal of Mathematical Psychology*, 42(2–3), pp.305–326.

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

Courchesne, E., Hillyard, S.A. & Galambos, R., 1975. Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology*.

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), pp.9–21.

Doll, B.B. et al., 2009. Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, 1299, pp.74–94.

Doñamayor, N. et al., 2011. Temporal dynamics of reward processing revealed by magnetoencephalography. *Human Brain Mapping*, 32(12), pp.2228–2240.

FABIANI, M. & FRIEDMAN, D., 1995. Changes in brain activity patterns in aging: The novelty oddball. *Psychophysiology*.

Friston, K.J. et al., 2017. Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, 77(April), pp.388–402. Available at: http://dx.doi.org/10.1016/j.neubiorev.2017.04.009.

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

Gl??scher, J. et al., 2010. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), pp.585–595.

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).

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, pp.948–959.

Keramati, M. et al., 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, 113(45), pp.12868–12873. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1609094113.

Kool, W., Gershman, S.J. & Cushman, F.A., 2017. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, 28(9), pp.1321–1333.

Kool, W., Gershman, S.J. & Cushman, F.A., 2018. Planning Complexity Registers as a Cost in Metacontrol. *Journal of Cognitive Neuroscience*, pp.1–14. Available at: https://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_01263.

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

Nieuwenhuis, S. et al., 2005. Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, 21(11), pp.3161–3168. Available at: http://doi.wiley.com/10.1111/j.1460-9568.2005.04152.x.

O'Doherty, J.P. et al., 2003. Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2), pp.329–337.

Opitz, B. et al., 1999. The functional neuroanatomy of novelty processing: Integrating ERP and fMRI results. *Cerebral Cortex*.

Otto, A.R. et al., 2013. The Curse of Planning. *Psychological Science*, 24(5), pp.751–761. Available at: http://journals.sagepub.com/doi/10.1177/0956797612463080.

Sutton, R.S. & Barto, A.G., 1998. Introduction to Reinforcement Learning. *Learning*, 4(1996), pp.1–5. Available at: http://dl.acm.org/citation.cfm?id=551283.

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), pp.1–11.

Tucker, D.M. et al., 2003. Frontolimbic Response to Negative Feedback in Clinical Depression. *Journal of Abnormal Psychology*, 112(4), pp.667–678.

Bibliography

Hirotugu Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control AC-19*, 19:716–723, 1974.

D. Alnaes, M. H. Sneve, T. Espeseth, T. Endestad, S. H. P. van de Pavert, and B. Laeng. 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(4), 2014.

Oscar Arias-Carrián, Maria Stamelou, Eric Murillo-Rodríguez, Manuel Menéndez-Gonzlez, and Ernst Pöppel. Dopaminergic reward system: A short integrative review. *International Archives of Medicine*, 3(1):1–6, 2010.

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

Gary Aston-Jones, Janusz Rajkowski, Piotr Kubiak, and Tatiana Alexinsky. Locus Coeruleus Neurons in Monkey Are Selectively Activated by Attended Cues in a Vigilance Task. *The Journal of Neuroscience*, 14(July):4467–4480, 1994.

Vivek R Athalye, Fernando J Santos, Jose M Carmena, and Rui M Costa. Evidence for a neural law of effect. *Science*, 359(6379):1024–1029, 2018.

Jan Balaguer, Hugo Spiers, Demis Hassabis, and Christopher Summerfield. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. *Neuron*, 90(4):893–903, 2016.

Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 1982.

Jeff A. Beeler. Thorndike's law 2.0: Dopamine and the regulation of thrift. *Frontiers in Neuroscience*, 6, 2012.

Timothy E. J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F.S. Rushworth. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–21, 2007.

Eduardo E Benarroch. The locus ceruleus norepinephrine system Functional organization and potential clinical significance. *Neurology*, 73(20):1699–1704, 2009.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author. *J. R. Statist. Soc.*, 57(1):289–300, 1995.

Anne Bergt, Anne E Urai, Tobias H Donner, and Lars Schwabe. Reading memory formation from the eyes. *bioRxiv*, page 268490, 2018. URL .

Gregory S Berns, Samuel M McClure, Giuseppe Pagnoni, and P Read Montague. Predictability modulates human brain response to reward. *J. of Neuroscience*, 21(8):2793–8, 2001. .

Kent C. Berridge. The debate over dopamines role in reward: The case for incentive salience. *Psychopharmacology*, 191(3):391–431, 2007.

Katie C Bittner, Aaron D. Milstein, Christine Grienberger, Sandro Romani, and Jeffrey C. Magee. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science*, 357(6355): 1033–1036, 2017.

Rafal Bogacz, Samuel M. McClure, Jian Li, Jonathan D Cohen, and P. Read Montague. Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, 1153 (1):111–121, 2007.

Aaron M. Bornstein and Kenneth A Norman. Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, 20(7):997–1003, 2017.

Matthew M. Botvinick, Yael Niv, and Andrew Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

Brainard D. H. The Psychophysics Toolbox. *Spatial Vision*, 10:433–436, 1997.

Michael Browning, Timothy E. J. Behrens, Gerhard Jocham, Jill X. O'Reilly, and Sonia J Bishop. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat Neurosci*, 18(4):590–596, 2015. .

Zuzanna Brzosko, Wolfram Schultz, and Ole Paulsen. Retroactive modulation of spike timingdependent plasticity by dopamine. *eLife*, 4:09685, 2015. .

Zuzanna Brzosko, Sara Zannone, Wolfram Schultz, Claudia Clopath, and Ole Paulsen. Sequential neuromodulation of hebbian plasticity offers mechanism for effective reward-based navigation. *eLife*, 6:27756, 2017.

Kenneth P. Burnham and David R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2):261–304, 2004.

Robert R Bush and Frederick Mosteller. A mathematical model for simple learning. *Psychological review*, 58(5):313, 1951.

Fabian Chersi and Neil Burgess. The Cognitive Architecture of Spatial Navigation: Hippocampal and Striatal Contributions. *Neuron*, 88(1):64–77, 2015. .

Anne G E Collins and Michael J Frank. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review*, 121(3):337–366, 2014. .

Matteo Colombo. Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 45(1):57–67, 2014. .

Nathaniel D. Daw and John P. O'Doherty. Multiple Systems for Value Learning. In Paul W. Glimcher and Ernst Fehr, editors, *Neuroeconomics: Decision Making and the Brain: Second Edition*, pages 393–410. Elsevier Inc., 2013. ISBN 9780124160088.

Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6): 1204–1215, 2011. .

Peter Dayan and Yael Niv. Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008. .

Peter Dayan and Angela J. Yu. Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4):335–350, 2006. URL .

Jan Willem de Gee, Olympia Colizoli, Niels A. Kloosterman, Tomas Knapen, Sander Nieuwenhuis, and Tobias H. Donner. Dynamic modulation of decision biases by brainstem arousal systems. *eLife*, 6:1–36, 2017.

Arnaud Delorme and Scott Makeig. EEGLAB: an open sorce toolbox for analysis of single-trail EEG dynamics including independent component anlaysis. *Journal of Neuroscience Methods*, 134:9–21, 2004. .

Bradley B. Doll, Katherine D Duncan, Dylan Alexander Simon, Daphna Shohamy, and Nathaniel D. Daw. Model-based choices involve prospective neural activity. *Nature neuroscience*, 18(5):767–772, 2015a.

Bradley B. Doll, Daphna Shohamy, and Nathaniel D. Daw. Multiple memory systems as substrates for multiple decision systems. *Neurobiology of Learning and Memory*, 117:4–13, 2015b.

Kenji Doya. Modulators of decision making. *Nature Neuroscience*, 11(4):410–416, apr 2008. ISSN 1097-6256. .

Katherine D Duncan and Daphna Shohamy. Memory States Influence Value-Based Decisions. 145(11):1420–1426, 2016. .

Wolfgang Einhäuser, Christof Koch, and Olivia L Carter. Pupil dilation betrays the timing of decisions. *Frontiers in human neuroscience*, 4(February):18, 2010. ISSN 1662-5161. .

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

Lotem Elber-Dorozko and Yonatan Loewenstein. Striatal Action-Value Neurons Reconsidered. *eLife*, 31(7), 2018. .

Eran Eldar, Valkyrie Felso, Jonathan D Cohen, and Yael Niv. A pupillary index of susceptibility to decision biases. *bioRxiv*, pages 1–18, 2018. .

Neir Eshel, Michael Bukwich, Vinod Rao, Vivian Hemmelder, Ju Tian, and Naoshige Uchida. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525(7568): 243–246, 2015. .

Simon D. Fisher, Paul B. Robertson, Melony J. Black, Peter Redgrave, Mark A. Sagar, Wickliffe C. Abraham, and John N.J. Reynolds. Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nature Communications*, 8(1), 2017. .

Cyrus K. Foroughi, Ciara Sibley, and Joseph T. Coyne. Pupil size as a measure of within-task learning. *Psychophysiology*, 54(10):1436–1443, 2017. .

Uwe Frey, Richard G M Morris, and Others. Synaptic tagging and long-term potentiation. *Nature*, 385(6616):533–536, 1997.

Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience, The Biology of the Mind*. W. W. Norton & Company, 3 edition, 2008. ISBN 978-0393927955.

Samuel J Gershman, Marie-H Monfils, Kenneth A Norman, and Yael Niv. The computational nature of memory modification. *eLife*, 6, 2017. URL .

Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 12:53, 2018. .

Jan Gläscher, Nathaniel D. Daw, Peter Dayan, and John P. O'Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010. .

Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108:15647–15654, 2011.

Paul W. Glimcher and Ernst Fehr, editors. *Neuroeconomics: Decision Making and the Brain: Second Edition*. Elsevier Inc., 2 edition, 2013. ISBN 9780124160088. .

Todd M. Gureckis and Bradley C. Love. Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, 113(3):293–313, 2009. .

Arif A. Hamid, Jeffrey R. Pettibone, Omar S. Mabrouk, Vaughn L. Hetrick, Robert Schmidt, Caitlin M. Vander Weele, Robert T. Kennedy, Brandon J. Aragona, and Joshua D. Berke. Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, 19(1):117–126, 2015.

W. K. Hastings. Monte Carlo simulation methods using Markov Chains and their applications. *Biometrika*, 57:97–109, 1970.

Kaiwen He, Marco Huertas, Su Z. Hong, Xiao Xiu Tie, Johannes W. Hell, Harel Shouval, and Alfredo Kirkwood. Distinct Eligibility Traces for LTP and LTD in Cortical Synapses. *Neuron*, 88(3):528–538, 2015.

Eckhard H. Hess and James M. Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

Bert Hoeks and Willem J M Levelt. Pupillary dilation as a measure of attention: a quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, 25(1):16–26, 1993. .

C. B. Holroyd and Michael G H Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4): 679–709, 2002. .

Jon C. Horvitz, Tripp Stewart, and Barry L. Jacobs. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research*, 759(2):251–258, 1997.

Eugene M Izhikevich. *Dynamical systems in neuroscience : the geometry of excitability and bursting*. MIT Press, 2007.

Gerhard Jocham, Kay H. Brodersen, Alexandra O. Constantinescu, Martin C. Kahn, Angela M. Ianni, Mark E. Walton, Matthew F.S. Rushworth, and Timothy E. J. Behrens. Reward-Guided Learning with and without Causal Attribution. *Neuron*, 90(1):177–190, 2016. .

Siddhartha Joshi, Yin Li, Rishi M. Kalwani, and Joshua I. Gold. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, 89(1):221–234, 2016. .

D Kahneman and J Beatty. Pupil diameter and load on memory. *Science (New York, N.Y.)*, 154 (3756):1583–5, 1966.

A Harry Klopf. Brain function and adaptive systems: a heterostatic theory, 1972.

Michal T. Kucewicz, Jaromir Dolezal, Vaclav Kremen, Brent M. Berry, Laura R. Miller, Abigail L. Magee, Vratislav Fabian, and Gregory A. Worrell. Pupil size reflects successful encoding and recall of memory in humans. *Scientific Reports*, 8(1):4949, 2018. .

Bruno Laeng, Sylvain Sirois, and Gustaf Gredebäck. Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7(1):18–27, 2012. .

Rylan S. Larsen and Jack Waters. Neuromodulatory Correlates of Pupil Dilation. *Frontiers in Neural Circuits*, 12(March):1–9, 2018. .

Claudio Lavín, René San Martín, and Eduardo Rosales Jubal. Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience*, 7(January): 1–8, 2014. .

M. Lengyel and Peter Dayan. Hippocampal Contributions to Control The Third Way. *Advances in Neural Information Processing Systems*, pages 1–8, 2008.

John Lisman. The Challenge of Understanding the Brain: Where We Stand in 2015. *Neuron*, 86 (4):864–882, 2015. .

Otto Lowenstein, Richard Feinberg, and Ie Irene E. Loewenfeld. Pupillary Movements During Acute and Chronic Fatigue A New Test for the Objective Evaluation of Tiredness. *Investigative Ophthalmology*, 2(1):138–158, 1963.

Adam Henry Marblestone, Greg Wayne, and Konrad P Kording. Towards an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10(10):94, 2016.

Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. Safe and sensible baseline correction of pupil-size data. *PeerJ PrePrints*, pages 1–25, 2017. .

Samuel M. McClure, Michele K. York, and P. Read Montague. The neural substrates of reward processing in humans: The modern role of fMRI. *Neuroscientist*, 10(3):260–268, 2004.

William Menegas, Benedicte M Babayan, Naoshige Uchida, and Mitsuko Watabe-uchida. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum. *eLife*, (6):21886, 2017. .

Marvin Minsky. Steps toward Artificial Intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961. .

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016. URL .

P.R. Montague, P. Dayan, and T.J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16:1936–1947, 1996.

Peter R. Murphy, Redmond G. O'Connell, Michael O'Sullivan, Ian H. Robertson, and Joshua H. Balsters. Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, 35(8):4140–4154, 2014.

Matthew R. Nassar, Katherine M Rumsey, Robert C. Wilson, Kinjan Parikh, Benjamin Heasly, and Joshua I. Gold. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, 15(7):1040–1046, 2012. .

A Newell. The chess machine: an example of dealing with a complex task by adaptation. *Proceedings of the March 1-3, 1955, western joint computer conference*, 4:101–108, 1955. .

Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3): 139–154, 2009.

Yael Niv, J. A. Edlund, Peter Dayan, and John P. O'Doherty. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, 32(2):551–562, 2012. .

J O'Doherty, P Dayan, K Friston, H Critchley, and R Dolan. Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during pavlovian appetitive learning. *Neuron*, 38:329–337, 2003.

John P. O'Doherty. Beyond simple reinforcement learning: The computational neurobiology of reward-learning and valuation. *European Journal of Neuroscience*, 35(7):987–990, 2012. .

John P. O'Doherty, Jeffrey Cockburn, and Wolfgang M Pauli. Learning, Reward, and Decision Making. *Annu. Rev. Psychol.*, 68:73–100, 2017. .

Samantha C. Otero, Brendan S. Weekes, and Samuel B. Hutton. Pupil size changes during recognition memory. *Psychophysiology*, 48(10):1346–1353, 2011.

G Pagnoni, CF Zink, PR Montague, and GS Berns. Activity in human ventral striatum locked to errors in reward prediction. *Nature Neuroscience*, 5(2):97–98, 2002.

Elise Payzan-LeNestour, Le Nestour, and Peter Bossaerts. Decision Making In Nonstationary Environments : An Experimental Study Of Human Adaptation. 2009.

Jing Peng and Ronald J Williams. Incremental Multi-Step Q-Learning. *Machine Learning*, 22: 283–290, 1996.

Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J. Dolan, and Chris D. Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045, 2006.

Kerstin Preuschoff, Bernard Marius 't Hart, and Wolfgang Einhäuser. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, 5: 1–12, 2011.

Antonio Rangel, Colin F Camerer, and P. Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556, 2008. .

P. Redgrave and K. Gurney. The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.*, 7:967–975, 2006.

Roger L. Redondo and Richard G. M. Morris. Making memories last: the synaptic tagging and capture hypothesis. *Nat Rev Neurosci*, 12:17–30, 2011.

Robert A Rescorla, Allan R Wagner, and Others. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99, 1972.

José J F Ribas-Fernandes, Alec Solway, Carlos Diuk, Joseph T. McGuire, Andrew Barto, Yael Niv, and Matthew M. Botvinick. A Neural Signature of Hierarchical Reinforcement Learning. *Neuron*, 71(2):370–379, 2011. .

Nina Rouhani, Kenneth A. Norman, and Yael Niv. Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(9):1430–1443, 2018. .

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, England, 1994.

R Salakhutdinov and J Tenenbaum. One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 195–207, 2012.

John D. Salamone and Mercè Correa. The Mysterious Motivational Functions of Mesolimbic Dopamine. *Neuron*, 76(3):470–485, 2012. .

K Samejima, Y Ueda, K Doya, and M Kimura. Representation of action-specific reward value in the striatum. *Science*, 310:1337–1340, 2005.

Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

E. Samuels and E. Szabadi. Functional Neuroanatomy of the Noradrenergic Locus Coeruleus: Its Roles in the Regulation of Arousal and Autonomic Function Part I: Principles of Functional Organisation. *Current Neuropharmacology*, 6(3):235–253, 2008. .

T. Schonberg, Nathaniel D. Daw, D. Joel, and John P. O'Doherty. Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. *Journal of Neuroscience*, 27(47):12860–12867, 2007. .

W. Schultz, P. Dayan, and R.R. Montague. A neural substrate for prediction and reward. *Science*, 275:1593–1599, 1997.

Wolfram Schultz. Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, 18(1):23–32, 2016. .

Daphna Shohamy and Nathaniel D. Daw. Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, 5:85–90, 2015. .

Dylan Alexander Simon and Nathaniel D. Daw. Neural correlates of forward planning in a spatial decision task in humans. *The Journal of Neuroscience*, 31(14):5526–5539, 2011. .

H M Simpson and Shirley M Hale. Pupillary Changes during a Decision-Making Task. *Perceptual and Motor Skills*, 29(2):495–498, 1969.

Satinder Singh and Richard S. Sutton. Reinforcement Learning with replacing elibibility traces. *Machine Learning*, 22:123–158, 1996. .

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine Learning*, 38(3):287–308, 2000. .

Thomas A. Stalnaker, Nisha K. Cooch, and Geoffrey Schoenbaum. What the orbitofrontal cortex does not do. *Nature Neuroscience*, 18(5):620–627, 2015. .

Caleb E. Strait, Brianna J. Sleezer, and Benjamin Y. Hayden. Signatures of value comparison in ventral striatum neurons. *PLoS Biology*, 13(6):1–22, 2015. .

Christopher Summerfield and Konstantinos Tsetsos. Do humans make good decisions? *Trends in Cognitive Sciences*, 19(1):27–34, 2015.

J. P. Sutton, J. S. Beis, and L. E. H. Trainor. Hierarchical model of memory and memory loss. *J. Phys. A*, 21:4443–4454, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, (in progress) second edition, 2018.

Emilie C.J. Syed, Laura L. Grima, Peter J. Magill, Rafal Bogacz, Peter Brown, and Mark E. Walton. Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nature Neuroscience*, 19(1):34–36, 2015. .

Lung Hao Tai, A. Moses Lee, Nora Benavidez, Antonello Bonci, and Linda Wilbrecht. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience*, 15(9):1281–1289, 2012. .

Elisa M. Tartaglia, Aaron M. Clarke, and Michael H. Herzog. What to choose next? A paradigm for testing human sequential decision making. *Frontiers in Psychology*, 8:1–11, 2017.

Edward L Thorndike. *Animal intelligence.* New York: Macmillan, 1911.

Edward L Thorndike. A proof of the law of effect. *Science*, 77:173–175, 1933.

Nash Unsworth and Matthew K. Robison. Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychonomic Bulletin and Review*, 22(3):757–765, 2015. .

Anne E Urai, Anke Braun, and Tobias H. Donner. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, 8:1–11, 2017.

Matthew M. Walsh and John R. Anderson. Learning from delayed feedback: Neural responses in temporal credit assignment. *Cognitive, Affective and Behavioral Neuroscience*, 11(2): 131–143, 2011. .

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

Joseph Tao-yi Wang. Eye Pupil Dilation and Eye Tracking. In *A handbook of process tracing methods for decision research: A critical review and user's guide*, pages 185–204. Psychology Press New York, 2011. ISBN 978-1848728646.

Mitsuko Watabe-Uchida, Neir Eshel, and Naoshige Uchida. Neural Circuitry of Reward Prediction Error. *Annual Review of Neuroscience*, 40(1):373–394, 2017.

B. Wilhelm, H. Giedke, H. Lüdtke, E. Bittner, A. Hofmann, and H. Wilhelm. Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. *Journal of Sleep Research*, 10(1):1–7, 2001. .

R.J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Reinforcement Learning*, 8:229–256, 1992.

G Elliott Wimmer and Daphna Shohamy. Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. *Science (New York, N.Y.)*, 338(10):270–3, 2012.

K. Wimmer, D. Q. Nykamp, C. Constantinidis, and A. Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.*, 17(3):431, 2014.

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

Klaus Wunderlich, Peter Dayan, and Raymond J. Dolan. Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, 15(5):786–791, 2012.

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

MARCO LEHMANN

PhD Cand Neuroscience (EPFL 2018)
MSc Computer Science (EPFL 2005)

Chemin de la Chiésaz 5, CH-1024 Ecublens
marco.lehmann@epfl.ch | +41 77 463 53 72
www.linkedin.com/in/marcophilipplehmann

Education & Research

2014 – 2018 PhD Cand Laboratory of Computational Neuroscience, EPF Lausanne

- Interdisciplinary research in reinforcement learning in humans.
  - “Evidence for eligibility traces in human learning” Preprint: https://arxiv.org/abs/1707.04192
  - "Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules", W. Gerstner, M. Lehmann, V. Liakoni, D. Corneil J. Brea, Frontiers in Neural Circuits, 12:53, 2018

- Conference Posters:
  - Reinforcement Learning and Decision Making (RLDM), 2017, Ann Arbor, Michigan, USA: “Eligibility trace signatures in human behaviour, pupil dilation and EEG”
  - Computational and Systems Neuroscience (COSYNE), 2017, Salt Lake City, USA: “Human learning in complex environments: episodic memory challenges the model-free–model-based realm.”
  - Computational and Systems Neuroscience (COSYNE), 2015, Salt Lake City, USA: “Bayesian filtering, parallel hypotheses and uncertainty: A new, combined model for human learning.”

- Conference Talk:
  - Lémanic Neuroscience Meeting (LNAM), 2017, Les Diablerets, CH: “Identifying distinct learning strategies in humans during a complex task”

- Teaching Assistant for Bachelor and Master classes:
  - “Programming in C++”, “Supervised and Unsupervised Learning”
  - “Biological Modelling of Neural Networks”. We developed a Python framework for this class which supports simulation, analyzation and visualization of neuronal dynamical systems: http://neuronaldynamics-exercises.readthedocs.io/en/latest/exercises/spatial-working-memory.html

- Bachelor & Master Project Supervision (selection):
  - «State Space Discretization Techniques for Reinforcement Learning»
  - «Storage and Recall of Correlated Sequences in Hopfield Networks»
  - «Surprise: from Theory to Experimental Design»

2008 CAS in Project Management (Zurich University of Applied Sciences, ZHAW)

2000 – 2005 MSc. Computer Science EPFL

- For my master project «Réalisation d’un serveur de corrections GPS accessible par GPRS» I was awarded the Prix UNICIBLE (for exceptional quality) and the Prix IGSO (for innovation).
- 2003/04: Exchange at the Royal Institute of Technology (KTH), Stockholm, Sweden.

1992 – 2000 Apprenticeship and “Zweitwegmatura”

- Apprenticeship as a land surveyor, Loser & Eugster, Gossau SG (1996)
- Interstaatliche Maturitätsschule für Erwachsene ISME, St. Gallen (2000)

Work Experience

2009 – 2013 Head of Software Development (GEOINFO, Herisau)

- Team Lead.
- Responsible for system architecture and successful service infrastructure upscaling.
- Classic and agile project management.

2005 – 2009 Software Engineer (GEOINFO, Herisau)

- Development of a Geographic Information Systems (GIS) (www.geoportal.ch)
- Responsible for service layer (SOA), database, data integration and analytics.
- Contributing to data processing pipeline: integration of heterogenous data sources, quality control, aggregation and pre-processing.

2005 – 2009 part-time Lecturer in Informatics

- Teaching «Object oriented programming in Java» and «Algorithms & Data structures» at the “Höhere Fachschule Uster” (HFU, http://www.hbu.ch/de/Informatik)

Skills & Technologies

Leadership & Management

- Change management: introduced SCRUM at GEOINFO. (Role: Scrum Master).
- Project management: supervised software projects with respect to time, quality and cost.
- Communication: speaker at events for employees, clients and public authorities.

Machine Learning & Modelling

- Experience in Deep- & Reinforcement Learning (Neuro & Machine)
- Classification | Dimensionality Reduction | Regression | Time-Series Analysis
- Bayesian Statistics | Dynamical Systems | Signal Processing

Programming Languages & Tools

- Languages: Python | Julia |Java | C# Dot.Net| Matlab | PL/SQL | Jupyter | SciPy
- Enterprise Application Architecture | Distributed Systems | RESTful API design
- ML and cloud: TensorFlow & Keras on Amazon AWS

Network & Communication

Junior Chamber International (JCI) (a global network of young leaders)

- 2010 – 2012 Committee member of the local chapter JCI Wil (SG), President 2011.
- Participation in leadership and coaching trainings.

Communication skills

- Public speaking: “Understanding the Brain: Machine Learning meets Neuroscience”. https://www.meetup.com/Swisscom-Digital-Lab/events/236631937/
- Passing complex ideas: lecturer for algorithms & data structures.
- Fluent written and spoken communication in: French (living in Lausanne for more than 8 years), English (scientific talks, business fluent), and German (native).