Preprint accepted for publication in *Neural Computation (2012)*

Solving the distal reward problem with rare correlations

Andrea Soltoggio[1] and **Jochen J. Steil**[1]

[1]Research Institute for Cognition and Robotics (CoR-Lab) and Faculty of Technology, Bielefeld University, Germany.

Keywords: Classical Conditioning, Instrumental Conditioning, Reward, Hebbian Plasticity, Neuromodulation.

Abstract

When learning by trial and error, the results of actions, manifested as rewards or punishments, occur often seconds after the actions that caused them. How can a reward be associated with an earlier action when the neural activity that caused that action is no longer present in the network? This problem is referred to as the *distal reward problem*. A recent computational study proposes a solution using modulated plasticity with spiking neurons, and argues that precise firing patterns in the millisecond range are essential for such a solution. In contrast, the current study shows that it is the rarity of correlating neural activity, and not the spike timing, that allows the network to solve the distal reward problem. In this study, rare correlations are detected in a standard rate-based computational model by means of a threshold-augmented Hebbian rule. The novel modulated plasticity rule allows a randomly connected network to learn in classical and instrumental conditioning scenarios with delayed rewards. The rarity of correlations is shown to be a pivotal factor in the learning and in handling various delays of the reward. This study additionally suggests the hypothesis that short-term synaptic plasticity may implement eligibility traces and thereby serve as a selection mechanism in promoting candidate synapses for long-term storage.

1 Introduction

Reward learning is a type of learning in which the causal relation between cues, actions and rewards is discovered. In classical conditioning (Pavlov, 1927; Skinner, 1953), a predicting stimulus becomes associated with the following reward or punishment. For example, in Pavlov's experiment, a dog associates the ringing of a bell with the subsequent food delivery such that, after a few trials, the dog starts salivating at the ring of the bell before the food comes. In instrumental conditioning (Thorndike, 1911;

Staddon, 1983), not only cues but also actions are associated with rewards. For example, in Thorndike's experiment, a cat becomes faster at escaping from a maze after learning the best route in a few attempts.

In both classical and instrumental conditioning, the rewards (or the results of actions) occur typically with some delay. Therefore, the correct association between cues (or actions), and rewards can be established only if a memory of what happened previously is maintained at the moment of reward delivery. Moreover, other disturbing cues or actions may intervene in between the predicting cues/actions and the reward. The problem of linking the correct past cues and actions to the subsequent rewards was named *distal reward problem* (Hull, 1943). How neural systems retain the memory of cues and actions, and how reward information is later processed to establish the correct associations is an important and open subject of investigation.

Reward stimuli in nervous systems have been related in the past two decades to changes in the activity of modulatory chemicals (Harris-Warrick and Marder, 1991; Marder, 1996; Marder and Thirumalai, 2002; Hasselmo, 2005; Rolls, 2009), particularly dopamine (Schultz et al., 1993; Wise, 2004; Arbuthnott and Wickens, 2007). Modulatory chemicals such as dopamine or acetylcholine (Bear et al., 2005), as the term *modulatory* indicates, are understood functionally not to affect excitation or inhibition strongly, as do glutamate and GABA, but rather to affect or modulate neural transmission and plasticity (Hasselmo, 1995; Bailey et al., 2000). Modulatory systems display unusually high or low activity in situations characterised by rewards, surprise or disappointment (Redgrave et al., 2008). The general idea is that an unexpected reward causes an increase in modulatory activity, while the failure to obtain an expected reward causes a depression of modulatory activity, which is also known to cause behaviour reversal (Deco and Rolls, 2005; O'Doherty et al., 2001). In fact, the activity of the modulator dopamine has been found to have similarities with the error signal in Temporal Difference (TD) learning (Schultz, 2002; Sutton and Barto, 1998; Rolls et al., 2008). Other neurotransmitters like acetylcholine, norepinephrine and serotonin were found to modulate a large variety of neural functions (Hasselmo, 1995; Marder, 1996; Bear et al., 2005). For example, the modulator acetylcholine was shown to enhance and stabilise learning and memory (for a review of related studies see Bailey et al. (2000)), revealing the central role of modulation in regulating long- and short-term plasticity (Kandel and Tauc, 1965). These findings provided inspiration for neural models that use neuromodulation to gate neural plasticity such that high modulation reinforces actions, while low modulation extinguishes actions (Abbott, 1990; Montague et al., 1995; Fellous and Linster, 1998; Porr and Wörgötter, 2007; Alexander and Sporns, 2002; Soula et al., 2005; Florian, 2007; Pfeiffer et al., 2010; Soltoggio and Stanley, 2012). Modulatory neurons were also shown to appear spontaneously in the evolution of artificial neural networks as evolutionarily advantageous traits in learning and memory tasks (Soltoggio et al., 2008). Whereas those studies show the advantage of increasing or decreasing neural weights based on modulatory signals, they do not address specifically the problem of accounting for delayed rewards. In real scenarios, rewards are often delayed, and other stimuli/actions intervene in between, making it difficult to establish which actions actually caused the reward. How neural systems solve the distal reward problem is not yet completely understood.

The increasing evidence from biology that modulation mediates reward learning has lead in recent years to a number of studies on reward-modulated spike-timing-dependent plasticity (STDP) and its learning properties (Soula et al., 2005; Florian, 2007; Farries and Fairhall, 2007; Legenstein et al., 2008; Potjans et al., 2009; Vasilaki et al., 2009; Potjans et al., 2011). In a seminal study, Izhikevich (2007) proposes a neural model to solve the distal reward problem using neuromodulated plasticity and spiking neurons. He suggests that the precise and coincident firing of spiking neurons generates local, chemical traces, named *eligibility traces* (Wang et al., 2000; Sarkisov and Wang, 2008), or *synaptic tags* (Frey and Morris, 1997; Redondo and Morris, 2011). The slow decay of traces helps reconstruct which synaptic connections were active previously in time when a delayed reward is delivered. The evidence and utility of synaptic tags in memory consolidation (Barco et al., 2008; Redondo and Morris, 2011) and in the solution of the distal reward problem (Päpper et al., 2011) has grown particularly in the last decade.

In Izhikevich (2007), although the spiking dynamics are in the millisecond scale, a reward delayed by seconds could correctly reinforce the responsible synapses. A randomly connected network of $1\,000$ neurons was shown to implement classical and instrumental conditioning, as well as learn to anticipate a reward delivery from an unconditioned stimulus to two predicting stimuli (conditioned stimuli). The results were attributed to the combination of the fast millisecond scale of spiking dynamics with the slower decay of eligibility traces. Izhikevich (2007) clearly points out that the precise spike timing is essential for STDP (Markram et al., 1997; Bi and Poo, 1998, 2001; Pawlak et al., 2010) to generate eligibility traces, and remarks how rate-based models cannot implement such dynamics.

In contrast to the previous studies cited above, the current work shows that precise spike timing is not the essential mechanism to implement the neural learning of classical and instrumental conditioning with delayed rewards. To support this claim, qualitatively identical learning dynamics to those in Izhikevich (2007) are reproduced with a rate-based model. The key computational principle in the present work is identified not in the spiking dynamics, but rather in the rarity of correlating neural activity, which in turn generates rare eligibility traces. The hypothesis is that a delayed reward can be correctly associated with earlier activity if the eligible synapses are a small percentage of the total number, regardless of whether the model is spiking or rate-based. Therefore, while Izhikevich (2007) uses spikes to generate rare eligibility traces, in the current study rare correlations are detected with a basic rate-based Hebbian rule augmented with a threshold. Such a threshold filters out the majority of correlating activity in the network. A small percentage of correlations generates eligibility traces. Those traces are modulated by reward to implement neural reward learning.

The novel Hebbian plasticity rule, named Rarely Correlating Hebbian Plasticity (RCHP), is tested first with simulations in four scenarios of classical and instrumental conditioning. The setup reproduces that used in Izhikevich (2007), except that here spiking neurons are replaced with rate-based neurons, and STDP with the new RCHP rule. The strikingly similar results produced by the rate-based model as compared to the spiking model lead to two main observations. First, millisecond-scale spiking neurons are not a computational requirement in this particular problem domain and can be replaced with rate-based neurons to implement similar weight change and learning. In the particular case of this study, the update frequency of neurons can be varied from the millisecond scale to the second scale, with an increase in computational efficiency up to three orders of magnitude. Such an improvement can be drastically beneficial in robotics and real-time applications. A second observation is that the rarity of correlations, rather than the spiking dynamics, is the essential property in both spiking and rate-based models that must be maintained to solve the distal reward problem. Such a position invites to revise current assumptions on the properties of rate-based Hebbian dynamics as opposed to STDP (Gerstner and Kistler, 2002; Cooper, 2005).

After establishing the equivalence of spiking and rate-based models in the solution of the distal reward problem, the study investigates further the pivotal role of rare correlations. One first aspect is that the rarity of correlations is functional for maintaining few eligible synapses in the network. By reducing further the rate at which traces are generated, traces can have longer decays. This experiment establishes for the first time a criterion for coping with different and long delays of rewards. Moreover, it allows for predictions regarding learning neural dynamics, which appear to be regulated by a balance between eligible synapses and rewarding episodes. A second experiment proves the independence of rare correlations from the particular way traces are implemented. In particular, connection weights themselves are shown to act as eligibility traces when the weights incorporate a long- and a short-term component. The short-term weight component replaces the synapse-specific chemical that represents the traces in Izhikevich (2007) and suggests an appealing new interpretation of the role of short-term plasticity. Finally, the important role of different rates and modalities of decay in traces and modulation is shown, further defining the relation between rare correlations and eligibility traces. In short, the principle of rare correlations empowers the basic Hebbian rule with unexpectedly effective learning dynamics with potentially large implications in the fields of neural plasticity and models of learning. This work, although it draws inspiration from Izhikevich (2007), is a fundamentally new study of the principle of rare correlations.

In the following section, the properties of a spiking model and STDP-modulated learning (as in Izhikevich (2007)) are briefly explained. Following, the new Rarely Correlating Hebbian Plasticity (RCHP) rule is introduced. The experimental setup is outlined in Section 3. The results in Section 4 reproduce four fundamental instances of classical and instrumental conditioning, which are used in this study, as well as in Izhikevich (2007), to test the ability of the network to solve the distal reward problem. Section 5 further investigates the learning with rare correlations, including the relation between the rarity of correlations and the rate of decay of traces, and the use of short-term plasticity as an eligibility trace. Additional experiments investigate the effect of various decays of the modulation and of the traces. The paper reports final considerations in the conclusion.

2 From spikes to rare correlations in rate coding

This section explores the role of rare neural correlations in the solution of the distal reward problem. The significance and utility of rare correlations is, in fact, the novel and main contribution of this study. Firstly, background information is provided with the description of the spiking neuron model used in Izhikevich (2007). The hypothesis is that learning in the spiking model may be fundamentally driven by rare correlations, and not by the precise timing of spikes. Following, the new Rarely Correlating Hebbian Plasticity (RCHP) rule, which extracts rare correlations from a rate coding, is introduced and explained. The main claim is that learning in conditioning scenarios is governed and regulated by the rarity of correlations, which can be equally represented in a spiking or rate-based model. The claim is demonstrated with extensive computer simulations presented in later sections.

2.1 The distal reward problem with a spiking model

In Izhikevich (2007), a network of $1\,000$ spiking neurons has 800 excitatory neurons and 200 inhibitory neurons randomly connected with 10% probability. This results in $100\,000$ synapses, $80\,000$ of which are plastic excitatory connections and $20\,000$ are fixed inhibitory connections. Neurons emit spikes at random times with a frequency of 1 Hz. Given two neurons i and j, such that j connects to i, every few minutes i fires spontaneously within 50 ms after j has fired. Such coincidence activates spike-timing-dependent plasticity (STDP), which in the cited study changes a synapse-specific value called *eligibility trace*, rather than the synaptic strength. The eligibility traces, one for each synapse, represent a chemical concentration c_{ji} at the synapse level between two neurons j and i. Traces have a decay time-constant τ_c of 1 s (Izhikevich, 2007, Fig.1c). The overall change of c_{ji} is expressed by

$$\dot{c}_{ji} = -c_{ji}/\tau_c + \mathrm{STDP_{ji}} \quad . \tag{1}$$

The eligibility traces c_{ji} are used to update the synaptic weights w_{ji}

$$\dot{w}_{ji} = c_{ji} \cdot d \tag{2}$$

where d is a global modulatory signal that represents the concentration of dopamine. The global signal d affects all the synapses in the network. The baseline (tonic) concentration of dopamine is assumed to be low (0.01 μM/s), which is increased to 0.5 when a reward is delivered. Therefore, a weight does not increase significantly its strength when STDP occurs, but rather when the global modulatory signal multiplies the synapse-specific eligibility trace. These dynamics are shown graphically in Fig. 1, which reproduces an illustration of the principle put forth in Izhikevich (2007, Fig. 1a-c, page 2444).

In one first experiment in Izhikevich (2007), one randomly chosen synapse σ, connecting neuron j^* to neuron i^*, is randomly selected from the $80\,000$ excitatory synapses. Each time the neuron i^* fires within 50 ms after neuron j^*, a reward is delivered to the whole network with a delay ranging from 1 to 3 s. Any other pair ji of connected neurons that produces consecutive firing changes its eligibility trace according to STDP but triggers no reward, which is only caused by STDP on the connection σ between neurons j^* and i^*. Reward is expressed by d, a global modulatory signal representing dopamine, which multiplies the eligibility traces across the whole network as in Eq. 2. Izhikevich (2007) reports that after one hour of simulated time and an average of 40 reward episodes, the connection σ between j^* and i^* reaches its maximum value while all other connections have lower or minimal values. Therefore, the network discovers

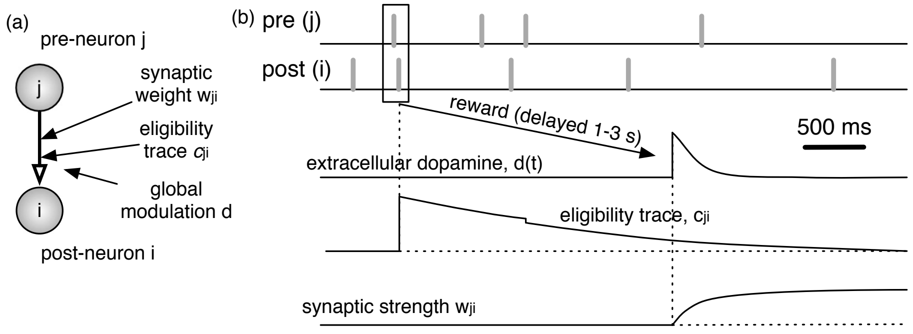

Figure 1: Graphical reproduction of the model in Izhikevich (2007, Fig. 1, page 2444) with spiking neurons and modulated eligibility traces. (a) Neurons j and i are connected with strength w_{ji}. Connections have their own eligibility traces c_{ji}. The level of global modulation d multiplies the eligibility traces to give the weight updates. (b) The presynaptic and postsynaptic neurons j and i fire at random times with an average frequency of 1 Hz. When neuron i fires shortly after neuron j, STDP increases the value of the eligibility trace c_{ji}, which then decays slowly over time. A subsequent reward signal d multiplies the eligibility trace to give the weight update. Therefore, the coincidence of spiking in the short 50 ms time window can be detected seconds later when the reward is delivered and causes a weight change.

the single pair of presynaptic and postsynaptic neurons (j^* and i^*), out of the $80\,000$ synapses, whose correlated firing causes a reward.

Interestingly, between the firing time of j^* and i^* and the time of reward (1 to 3 s later), the network continues to fire randomly and other eligibility traces are generated by coincident firing. Nevertheless, the network correctly identifies the unique synapse that triggers the reward. The network is therefore capable of associating a later reward with a previous event even when other events occur in between the triggering action and the delayed reward.

How can a network of a hundred thousand synapses discover the unique, randomly chosen synapse σ after only 40 reward episodes? Izhikevich (2007, page 2451, Section "Spiking versus mean Firing Rate Models") explains that the precise timing of the spikes allows the network to reconstruct the correct association between actions and rewards, while a rate-based model would fail. This notion is challenged in this study by using a rate-based model to generate the same learning of the spiking network. The present claim is that the essential computation to solve the distal reward problem is performed by the rarity of eligibility traces, and not by the spiking dynamics. Coincident spikes in Izhikevich (2007) are rare and therefore lead the network to maintain only few high eligibility traces at any time. When j^* and i^* fire in rapid succession, and a reward is delivered, some other connections in the network have high eligibility traces, but they are a small percentage of all connections in the network. The next time j^* and i^* correlate, and another reward is delivered, some other connections are eligible, but they are yet another small set of synapses. Therefore, while the trace of the connection σ is always significant when the reward is delivered, the other synapses are eligible only occasionally. For example, if 1% of synapses are eligible at any time, the first reward delivery reinforces approximately 800 synapses (1% of $80\,000$), together with the triggering synapse σ. The second reward delivery, after another correlating episode of σ, reinforces other 800 synapses, out of which on average only 8 (1% of 800) were reinforced before. After only three or four reward episodes, although modulation affects all synapses in the network, σ can be expected to be the only connection that was reinforced consistently.

According to this interpretation, the computational ground for solving the distal reward problem in Izhikevich (2007) does not lay in the spiking dynamics per se, but rather in the condition that correlating episodes, which generate traces, are rare. If so, one hypothesis is that the spiking dynamics in Izhikevich (2007) are used with the sole purpose of generating rare eligibility traces. Therefore, in contrast to the position in Izhikevich (2007, page 2451, Section "Spiking versus mean Firing Rate Models"), if traces were not rare, for example because of higher firing rates than 1 Hz, the spiking model would fail. By understanding the exact principle that implements the learning in the distal reward problem, a number of predictions on the learning dynamics can be formulated. Moreover, if rare correlations can be computed without spikes, the precise learning dynamics can be reproduced with a rate-based model. This hypothesis is tested by introducing a Hebbian rule that detects rare correlations.

2.2 Rarely Correlating Hebbian Plasticity

Hebbian plasticity for a rate-based discrete time network is expressed in its simplest form by

$$\Delta w_{ji}(t) = v_j(t - t_{pt}) \cdot v_i(t)\,, \tag{3}$$

where v_j is the output of a presynaptic neuron j, v_i the output of a postsynaptic neuron i, w_{ji} the strength of the connecting weight and t_{pt} is an interval corresponding to the *propagation time* for the presynaptic signal to reach the postsynaptic neuron. In the simplest implementation (also adopted in this study), t_{pt} is assumed to be one single step of computation, i.e. signals transfer from one neuron to another in one step. This setting represents the simplest way to implement the propagation delay that is necessary with recurrent connections and temporally-causal plasticity. The propagation delay is equal across all synaptic connections. This implies that any two connected neurons affect each other with a uniform delay of t_{pt} across the whole network. Longer or variable propagation times were proved to be useful for other types of computation, e.g. in polychronization (Izhikevich, 2006), but are not necessary in the present context and therefore are not further considered. The product in Eq. 3 describes how the presynaptic signal correlates with the postsynaptic signal after it has reached the postsynaptic neuron. The product attempts to measure how much the presynaptic neuron contributes to the activity of the postsynaptic neuron, implying a causality of activities rather than their simultaneity. Such a product returns a continuous value representing a *level* of correlation. In other words, all neurons correlate to a certain level. This is in contrast to STDP, in which the firing of two neurons triggers a response only if it occurs in a ± 50 ms time window. With neurons firing at an average frequency of 1 Hz, such an event is rare. However, rare correlations can be detected also in a rate-based model simply by introducing a threshold on the product of Eq. 3. Products exceeding such threshold return a positive value, while the others return zero. Similarly, a second threshold can be used to detect rare decorrelations, analogously to long-term depression (LTD) in STDP, which occurs when the postsynaptic neuron fires shortly before receiving a spike. This novel and yet simple rule is named Rarely Correlating Hebbian Plasticity (RCHP) and can be expressed as

$$\mathrm{RCHP}_{ji}(t) = \begin{cases} +0.5 & \text{if } v_j(t - t_{pt}) \cdot v_i(t) > \theta_{hi} \\ -1 & \text{if } v_j(t - t_{pt}) \cdot v_i(t) < \theta_{lo} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where θ_{hi} is the threshold value that determines when two rate-based neurons correlate strongly and θ_{lo} determines when two neurons decorrelate. The magnitude of the decrement triggered by decorrelation is twice that of the increase triggered by correlation to ensure that weights do not grow too large on average. Although other choices are possible, this study adopts such a setting to reproduce the asymmetry of the areas of long-term potentiation (LTP) versus long-term depression (LTD) in STDP (Magee and Johnston, 1997; Markram et al., 1997; Bi and Poo, 1998, 2001; Pawlak et al., 2010).

2.3 Detecting rare correlations

With spiking neurons, the probability of correlation or decorrelation depends on the width of the STDP time window, the frequency of spiking, synaptic strength, connectivity and internal neural dynamics; in a discrete time rate-based model, different factors must be considered. Among these are the range of weights, the range of the output, the transfer function, the intensity of neural noise and inputs. These factors are notoriously different from model to model, and like in the case of inputs and weight strengths, can change over time. Therefore, rather than devising analytically a fixed threshold, a more general approach is to estimate θ_{hi} and θ_{lo} during on-line simulation. The relevant aspect is to ensure that over a certain period of time the detected correlations are low. In the current experiments, θ_{hi} and θ_{lo} are estimated on-line to target 1% of correlations per second. Various approaches can be used. In the present experiments, the simulation starts by monitoring the correlation activity across the network and finds the threshold that results in the 1% rate for each second of simulation. The average threshold across samples is applied. After collecting 10 such samples (i.e. after 10 seconds), the network updates the 10-sample array only if the correlations per second exceed the target by $\pm 50\%$. Preliminary simulations showed that the target value of 1% of correlations is robust, allowing the algorithm to work reasonably well in the range $[0.1\%, 2\%]$. In the experiment section, the effect of varying the rarity of correlations is investigated in detail.

It is important to note that Eq. 4 per se does not represent the central contribution of this study, but it rather expresses one possible way to extract rare correlations (or causal events) from the network activity. Other types of mapping functions can be devised. The central aspect is instead the use of rare events that, as shown in the experiment sections, prove to be crucial in maintaining the correct balance of eligibility traces and consequently the correct learning dynamics. To illustrate the difference between plain Hebbian plasticity and the RCHP, Fig. 2 compares the outputs of the Hebbian rule and

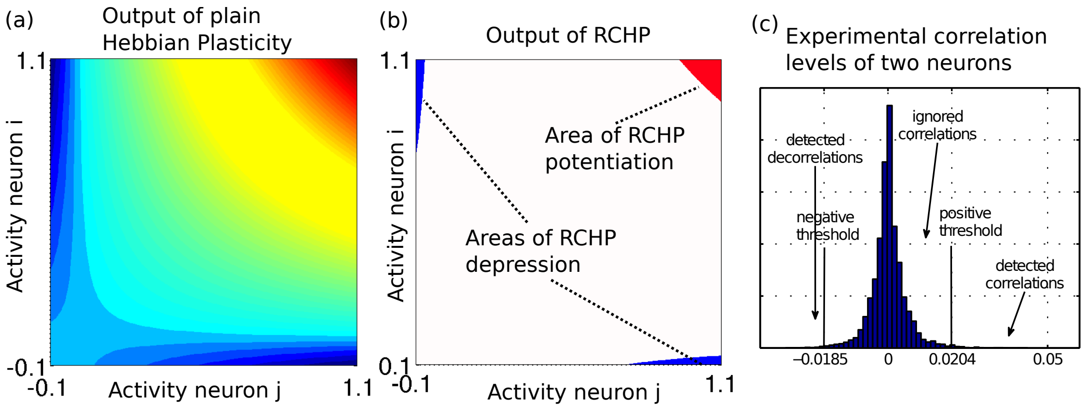

Figure 2: Outputs of the plain Hebbian rule and of the novel RCHP rule. (a) The temperature graph shows the possible products of two presynaptic and postsynaptic neuron activities in plain Hebbian plasticity as in Eq. 3 when the output varies in the interval $[-0.1, 1.1]$. (b) Output of the RCHP rule of Eq. 4. The narrow areas indicate rare decorrelations (1%) that cause depression. The small top-right area indicates rare correlations (1%) that cause potentiation. The large white area indicates that those levels of correlation are ignored by the RCHP rule, i.e. they neither cause changes in the eligibility traces nor in the weights. To extract the correct level of rare correlations, the thresholds are set according to the measured level of activity in a specific network as in (c), where the histogram shows experimentally measured correlation levels between two connected neurons during spontaneous activity in the model of the present study.

of the RCHP rule. Figs. 2b and c shown that 98% of correlating activity is ignored by the RCHP rule.

2.4 Varying the update frequency

Discrete time rate-based neurons are used instead of spiking neurons. All other network features, parameters and eligibility traces are as in Izhikevich (2007). The continuous time Eq. 1, when integrated in discrete time becomes

$$c_{ji}(t + \Delta t) = c_{ji}(t) \cdot e^{\frac{-\Delta t}{\tau_c}} + \mathrm{RCHP}_{ji}(t) \quad . \tag{5}$$

Similarly, Eq. 2 becomes

$$\Delta w_{ji}(t) = c_{ji}(t) \cdot d(t) \quad , \tag{6}$$

where $d(t)$ is the modulation level that is set to 0.12 per reward episode and 0 otherwise.

A central assumption in the RCHP rule is that it is the rarity of correlations, and not the precise spike timing, that is relevant to solve the distal reward problem. Therefore, an essential point is to show that the sampling time, or frequency of update, can be varied arbitrarily without affecting the learning dynamics. The integration time step Δt can be set according to the required speed of computation, determined mainly by the required input and output refresh rates. The interval Δt corresponds also to the propagation time t_{tp} in Eq. 4, implying that the internal speed of signal propagation can be as slow as the refresh rate of the input-output signals. As a consequence, and as opposed to Izhikevich (2007), the computation of the RCHP is based on the rarity of average correlations over time, and not on the precise timing of the neural dynamics. Sampling steps spanning over two orders of magnitudes (in the interval $[10, 1\,000]$ ms) are successfully tested in the simulations presented later in Section 4.

3 Experimental setup

The rate-based neural network in all experiments has 800 excitatory neurons and 200 inhibitory neurons. Each neuron has 100 random afferent connections, i.e. it has probability 0.1 of being connected with any other neuron. As in Izhikevich (2007), inhibitory neurons are one-fifth the total number of neurons but are five times more effective than excitatory neurons; their output is multiplied by a factor κ equal to -5 (the minus indicates inhibition), while κ for excitatory neurons is $+1$. The membrane potential u and output v for a neuron i are given by

$$u_i(t) = \sum_j (w_{ji} \cdot v_j(t) \cdot \kappa_j) \tag{7}$$

$$v_i(t + \Delta t) = \begin{cases} \tanh\big(\gamma \cdot u_i(t)\big) + \xi_i(t) & \text{if } u_i \geq 0 \\ \xi_i(t) & \text{if } u_i < 0 \end{cases} \tag{8}$$

where j is the index of a presynaptic neuron, γ is a constant gain parameter set to 0.2 and ξ_i is a noise signal drawn from a uniform distribution in the interval $[-0.15, 0.15]$. Excitatory weights vary in the range $[0, 1]$ and are initialised to small values in the range $[0, 0.01]$.

The neural noise ξ is an essential component of the neural computation to implement spontaneous activity. In fact, without noise and external inputs, all neuron outputs v and membrane potential u would be zero. In this case, all correlations are zero and it is therefore not possible to extract the 1% of connections with higher correlations. Neural noise can thus be interpreted as a generator of spontaneous activity.

Implementation details and the summary of the experimental settings are listed also in the Appendix 5.5.

4 Simulating the learning dynamics of the RCHP rule

Four established experiments of classical and instrumental conditioning are described and reproduced in this section. The purpose is to demonstrate that the proposed rate-based model with the Rarely Correlating Hebbian Plasticity (RCHP) solves these problems, reproduces the learning dynamics of the spiking model in Izhikevich (2007), and thereby proves that it is the rarity of correlations and not the precise firing pattern that solves the distal reward problem.

The spiking dynamics are substituted with the rate-based dynamics of Eq. 8. STDP is replaced by the newly introduced RCHP of Eq. 4. Eligibility traces have dynamics regulated by the decay and RCHP as prescribed by Eq. 5. All other settings are reproduced as in Izhikevich (2007) unless otherwise specified. Tables summarising the specific settings of each simulation are reported in the Appendix. The simulation algorithms and data in this study are reproducible with the Matlab scripts provided as part of this study and are available for downloaded at the author's associate website http://andrea.soltoggio.net/RCHP.

4.1 Reinforcing a synapse

The experiment described in this section shows how the RCHP rule can identify the unique synapse, out of $80\,000$, that is responsible for triggering a reward.

Two random excitatory neurons i^* and j^* are selected in the randomly connected network, such as j^* connects to i^*. Their connecting weight σ is set to zero at the beginning of the simulation. Each time the pair j^*i^* correlates according the RCHP rule of Eq. 4, i.e. each time $v_{j^*} \cdot v_{i^*}$ reaches the threshold θ_{hi}, a reward is delivered to the whole network with a variable delay ranging from 1 to 3 s. The time-constant τ_c is set to 2 s to give the traces a sufficient duration to cover the delay of the reward.

A property of the RCHP rule is that as long as correlations are rare during a unit of time, the sampling step can be freely chosen, thereby proving that the precise spike-timing is not essential in the solution of the problem. To prove this point, three simulations were conducted with sampling steps of 10 ms, 100 ms and 1 s, where 1 s was chosen as the longest sampling step to accommodate a reward delivery in the interval $[1, 3]$ s. To ensure similar reward rates across simulations with different time steps, a minimum interval between reward episodes is set to 6 s.

Figs. 3a-c show the results of the three simulations. At the beginning of the simulation, noise drives the activities of i^* and j^* to have a high correlation every few minutes. This causes the delivery of a delayed reward. With time and more rewards being delivered, the connection σ increases more than any other connection. In Figs. 3a-c, the central and right graphs show the histograms (in logarithmic and linear scale) of the connection strengths at the end of the simulation. In all three simulations, the second largest weight among the $80\,000$ is less than 50% of the weight σ, indicating the clear separation between the synapse σ and all of the others. The plots show that the weight dynamics do not change qualitatively when the frequency of computation is changed in the range from 10 ms to 1 s. This is a demonstration that the exact timing of network updates is not a crucial aspect in this type of computation.

To test the consistency of these results against the stochastic processes involved in the simulations, 120 runs (40 for each of the three chosen sampling steps) with different random seeds were launched, yielding similar learning dynamics. All plots are available for download as support material. In six cases out of 120, other weights beside σ grew to saturation. Interestingly, those other weights that grew to saturation are near σ. Their activity contributes to, or is caused by, the correlation of σ. In these cases, the network was able to trigger rewards at high rates, due to sustained activity around the synapse σ.

Additional tests, not shown in the present work, indicate that a number of network parameters affect the overall level of network activity. The balance between excitatory and inhibitory neurons is essential in maintaining the network activity low. The neural gain, the weight range and the density of connectivity determine the level to which each presynaptic neuron can drive a postsynaptic neuron. It is important to note that, in reward learning, the capability of growing functional pathways is essential. However, strong pathways reduce the randomness in the network, decrease decorrelations and increase correlations among connected neurons, both in spiking and rate-based models, potentially triggering a positive unstable feedback due to the Hebbian rule. In particular, the values for increments and decrements of the traces given in Eq. 4 (also in Izhikevich (2007)) indicate that each decorrelation episode (decrement -1) counts as two correlation episodes (increment +0.5). Therefore, under rewarding conditions, i.e. when the network receives frequent rewards, any two neurons that on average correlate more than twice the times they decorrelate are expected to grow their synaptic connection, regardless of whether they cause the reward or not. The network stability is therefore strongly related to maintaining very weak connections among neurons. The rate of weight update and the amount of the modulatory signal determine the rate of growth of weights. Those features of the network appear to work in unison with each other to maintain correct neural dynamics, plasticity and learning. Therefore, the principle of rare correlations guarantees the correct solution of the distal reward problem, but the network stability appears to depend also on a larger set of parameters.

Fig. 4a shows how the logarithmic histogram of the weight changes throughout time during the execution of a typical run. Occasionally, other synapses that do not trigger rewards increase temporarily, but the effect of decorrelations cause those weights to remain low in the long term. The percentage of high correlations per second during an arbitrarily chosen phase of the simulation is shown in Fig. 4b. The graph shows that the percentage of correlations varies but remains close to the 1% target, confirming that the network operates with a level of rare correlations.

The simulations in this section reproduce a simple form of instrumental conditioning in which the action of correlating neurons j^* and i^* is rewarded. The weight dynamics do not appear to be significantly different in the three simulations despite the use of three neuron-update frequencies separated by two orders of magnitude. In other words, in opposition to Izhikevich (2007), the precise timing of neural updates is not relevant for the learning dynamics and the solution of this experiment. Moreover, the fastest simulation with 1 s sampling time can be up to a thousand times computationally more efficient than a simulation with a millisecond sampling time. The weight update performed by the RCHP rule on rate-based discrete time neurons does not appear to be significantly different from the results of the spiking model as shown in the corresponding Fig. 1 in Izhikevich (2007, Page 2444).

4.2 Classical (Pavlovian) conditioning

In a second experiment, the capability of the network to perform classical conditioning is tested. One hundred possibly overlapping sets of neurons, $S_1, ..., S_{100}$, each set composed of 50 randomly chosen neurons, represent 100 different stimuli. A stimulus is imparted to the network by increasing by 20 the activation u of each neuron in the respective group. Stimuli are delivered to the network in a random sequence with an inter-stimulus delay in the interval $[100, 300]$ ms, therefore resulting in an average of five stimuli per second. When S_1 is delivered, a reward follows with a random delay up to 1 s as in Fig. 5a. Stimuli $S_2 ... S_{100}$ do not cause a reward. The delay between S_1 and the reward implies that other stimuli occur after S_1 and before the reward: this makes it more difficult for the network to discover which stimulus is actually responsible for the

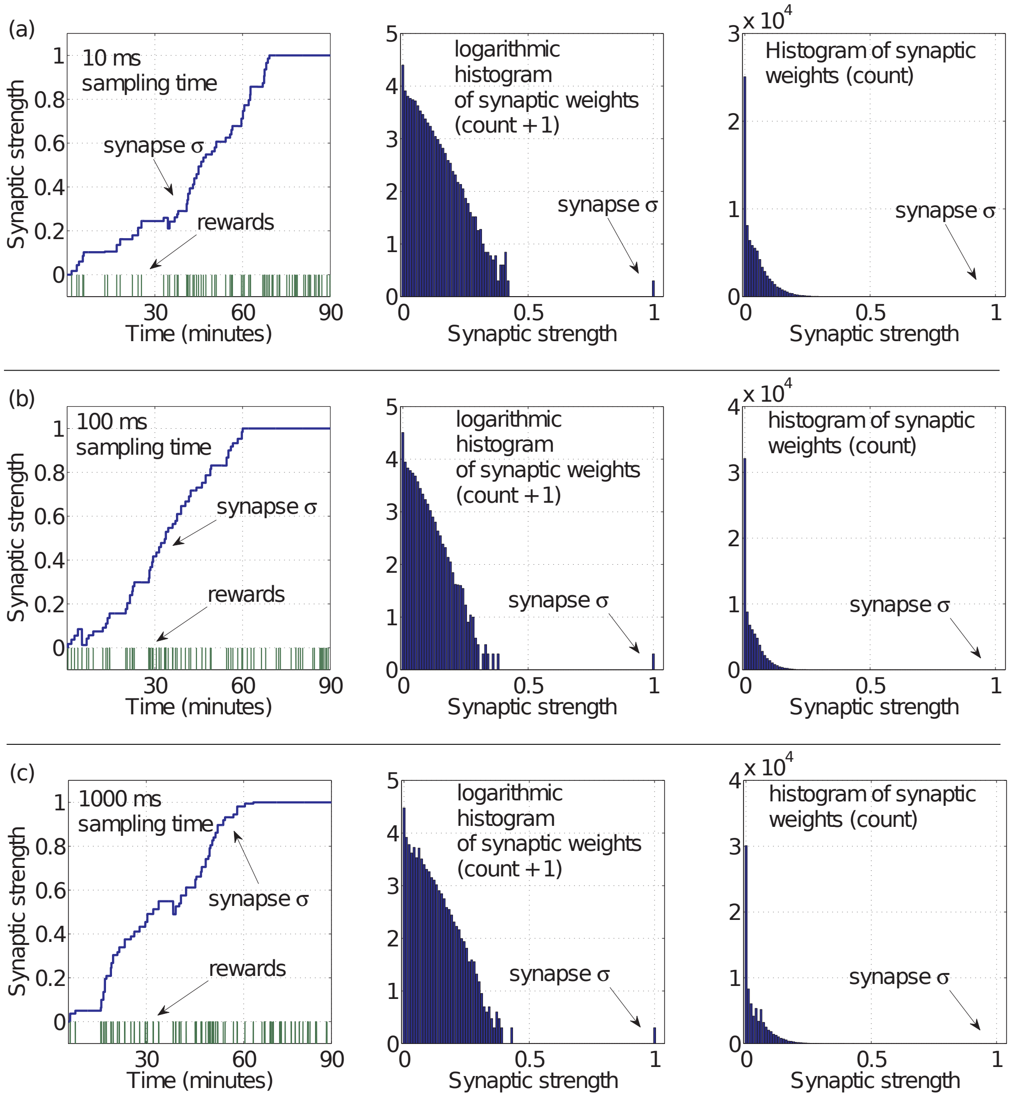

Figure 3: Instrumental conditioning of one synapse. (a) In the left graph, the strength of the synapses σ, initially set to zero, is shown to increase and reach the maximum value during the simulation with a sampling step of 10 ms. At the bottom of the graph, bars indicate the times of reward. The central figure shows the histogram of all connection strengths (weight binning 0.01, logarithmic scale), at the end of the simulation: only the synapse σ reaches the maximum value of 1. The right figure is the histogram of all connection weights in natural scale. (b) As in (a) with a 100 ms sampling step. (c) As in (a) and (b) with a $1\,000$ ms sampling step.

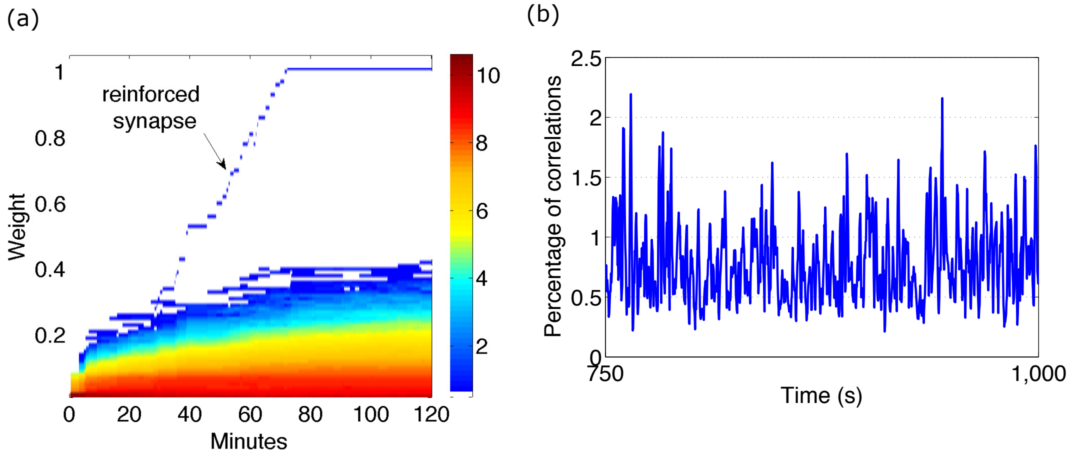

Figure 4: Analysis of instrumental conditioning of one synapses. (a) Logarithmic histogram of weights through time for a full run. The weight binning is 0.01. The colour code indicates the logarithmic value of the histogram, which is descriptive of the number of synapses that have a given weight. The reinforced synapse is observed to grow faster and separate itself from the other $79\,999$ synapses. (b) Percentage of correlations per second during an arbitrarily chosen 250-second simulation period. The plot shows that the percentage of correlations varies approximately around the target of 1%.

reward.

Initially, as shown in Fig. 5b, stimulus S_1 elicits a response in the network similar to other stimuli. However, with time the network starts to respond more to S_1. After one hour of simulated time, the network responds strongly to S_1 as shown in Fig. 5c. In other words, after some time the network starts to listen more to neurons in the S_1 group even though S_1 occurs with the same frequency as the other 99 stimuli. These learning dynamics are qualitatively identical to those in the spiking model in Izhikevich (2007, Fig. 2, page 2446).

An important question is, how can the network increase its response to S_1 out of 100 stimuli, all of which are equally likely to occur between S_1 and the reward? After each stimulus, some neurons in the network correlate with the stimulus' neurons and their connections increase their eligibility traces. The traces have a decay of 1 s (τ_c = 1) and therefore are significant in magnitude when the delayed reward occurs. The key factor in making the problem solvable is to have a large pool of stimuli, in this case 100, a few of which can randomly occur in between S_1 and the reward. Therefore, if one random stimulus, say S_{39} as in Fig. 5b, occurs in between S_1 and the reward, it is strengthened more than S_1 in that particular reward episode. However, the probability that S_{39} occurs again shortly after S_1 becomes lower as the set of possible stimuli becomes larger. If the pool of stimuli contained fewer stimuli, for example 5 instead of 100, those five stimuli would appear repeatedly within the reward time window, making it drastically more difficult to infer which is the reward-causing stimulus. In fact, if few stimuli occur continuously, or very frequently, they maintain high eligibility traces across synapses, thereby inducing facilitation when a reward occurs (data not shown). Therefore, the solution to this problem is found under the conditions that the *disturbing* stimuli are randomly selected from a large pool. Interestingly, if one disturbing stimulus occurs very often before the reward, even if it is not causally related to the reward delivery, the network would start to respond to it in a seemingly *superstitious* attitude. Such superstitious behaviours are frequently observed in conditioning experiments with animals (Timberlake and Lucas, 1985).

This experiment shows that the modulated RCHP rule with eligibility traces implements classical conditioning. The network becomes more reactive, i.e. it displays a higher neural activity, in response to the only stimulus that is associated with a reward, even if disturbing stimuli occur in between the conditioned stimulus S1 and the reward.

4.3 Stimulus response instrumental conditioning

This section considers a basic instance of instrumental conditioning in which the network is rewarded according to which action it performs, as in Izhikevich (2007, Problem 3, page 2447). Fifty neurons are chosen randomly to form a group S, where a stimulus is delivered. Two other non-overlapping groups of 50 random neurons, A and B, represent two groups of output neurons. Every 10 s a stimulus is delivered through S for 200ms. The output in groups A and B[1] is measured after the delivery of S. If $\|\mathrm{A}\| > \|\mathrm{B}\| + 1$, action A is performed; if $\|\mathrm{B}\| > \|\mathrm{A}\| + 1$, action B is performed. No action is performed if the difference between the outputs is less than 1. In a first simulation, when the network performs action A, a reward is given with a delay proportional to the difference $\|\mathrm{A}\| - \|\mathrm{B}\|$, so that a stronger response, expressed by a higher difference, results in a faster reward. No reward is given if action B, or no action, is performed. In a second simulation, action B is rewarded, but no reward occurs when action A, or no action, are performed.

Fig. 6a shows that when action A is rewarded, the network converges to select action A more frequently until B does not occur anymore. Conversely, when action B is rewarded, the network converges to perform action B more frequently (Fig. 6b). The results of this experiment differ slightly from those in Izhikevich (2007) in which the network could switch from action A to action B during the same simulation when the reward policy was changed. Preliminary experiments showed that the current network cannot switch after it has learnt to perform one action. However, the stability of response, i.e. the consistency in performing one action, appears to be greater in the present experiment than it was in Izhikevich (2007, Fig. 3, Page 2448) and could explain why the network cannot reverse its preference: the extinguished action becomes very unlikely after a certain period of learning. When reward conditions are switched, the extinguished action simply does not occur and rewards are not triggered. One possibility to reverse actions even in the current study is to deliver a negative reward, or punishment, to extinguish the action that was learnt and increase the probability of the complementary action (Deco and Rolls, 2005).

The two parallel simulations, one in which action A is rewarded and the other in which action B is rewarded, show that when convergence to the rewarding action is particularly fast in one case, it tends to be particularly slow in the complementary case. For

[1] Measured as $\|\mathrm{A}\| = \sum^{k\in\mathrm{A}}(v_k)$ and similarly for B.

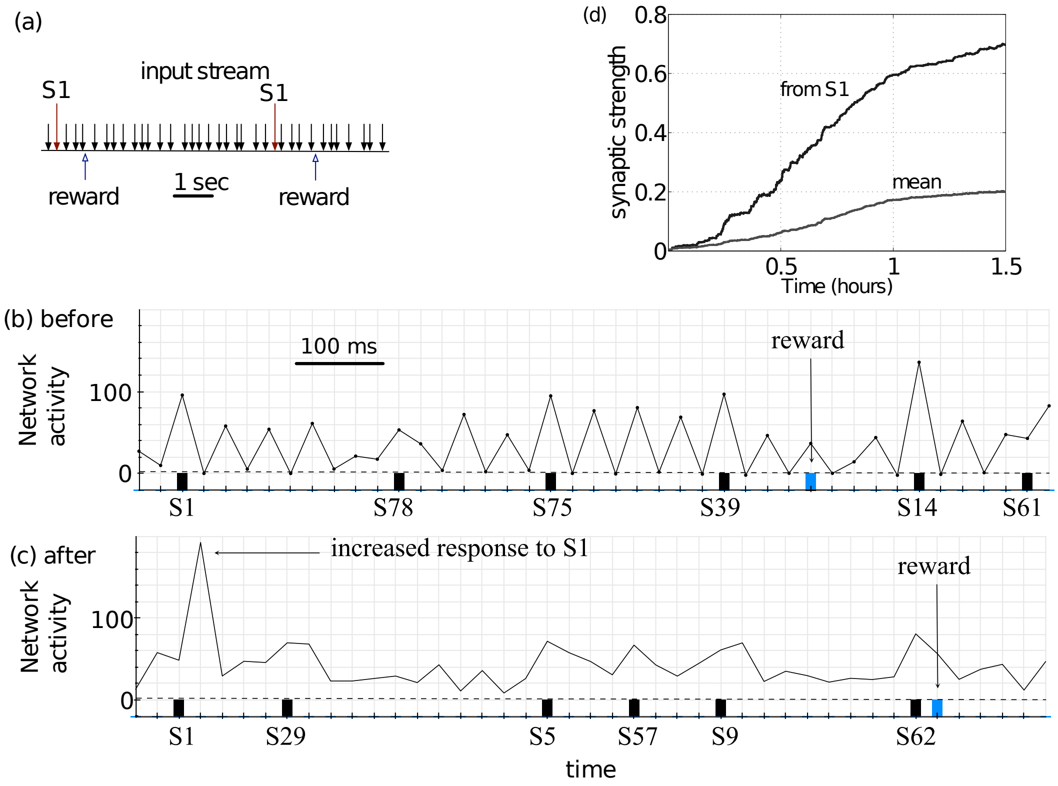

Figure 5: Classical (Pavlovian) Conditioning. (a) The stream of inputs is a random sequence of stimuli $S_i, i \in [1, 100]$ separated by random intervals between 100 ms and 300 ms. After an occurrence of S_1, a reward is delivered with a random delay up to 1 s. (b) Activity of the network (sum of all neuron outputs) at the beginning of the simulation: the stimulus S_1 elicits a small response from the network, similarly to the response of other stimuli. (c) After one hour of simulated time, stimulus S_1 elicits a strong response from the network. The stimulus has been associated with the reward and it is now generating a stronger response. (d) The average connection strength from S_1 grows to more than three times than of the other connections. This implies that the network associates S_1 with the reward by increasing the connection strength from neurons in S_1 to other neurons in the network.

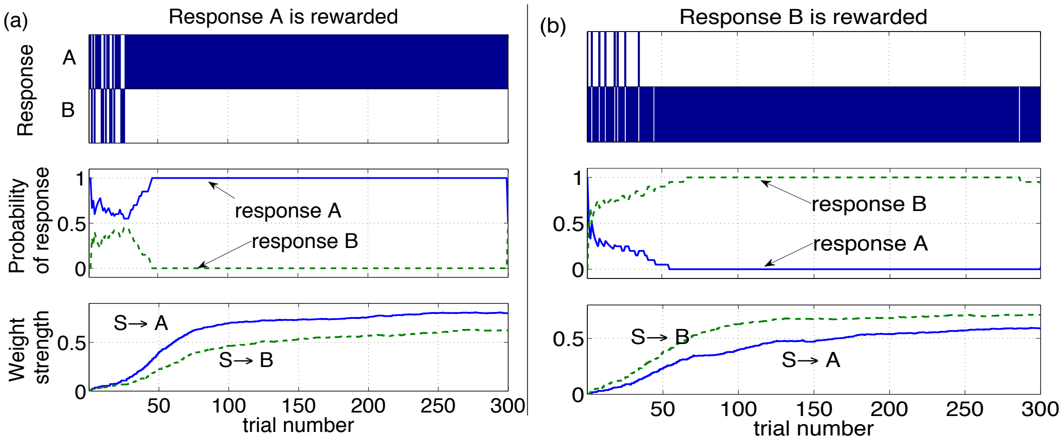

Figure 6: Instrumental Conditioning. (a) In a first simulation the action A is rewarded. The network initially explores both actions A and B but after only 30 trials converges to action A. The probability of A, computed as the frequency over the previous 20 trials, becomes 1 while the probability of B becomes 0. The average strength of connections from S to A emerges to be stronger than those from S to B. (b) In a second simulation that started with the same network (i.e. same initialisation of random weights), action B is rewarded. The network converges to choosing constantly action B in less than 50 trials. The probability of choosing action B converges to 1. Accordingly, the average strength of connections from S to B is higher than S to A. The plots indicate that the network can learn quickly which action leads to a reward.

example, in Fig. 6a, the network learns to perform action A with less trials than when it learns action B in Fig. 6b. The consistency of this observation across different random initial networks suggests that these learning dynamics are not completely independent from initial conditions such as the random connectivity.

It is also important to note that, although Izhikevich (2007) points out the high number of ways (10^{164}) in which groups A and B can be randomly chosen out of 800 neurons, a simpler consideration suggests that what matters is simply whether $\|A\|$ is greater than $\|B\|$ or not. For example, if A is the rewarding action, and the network choses A (by having higher activity in group A), the connections S-to-A have higher eligibility traces than the connections S-to-B. Therefore, the subsequent reward reinforces the S-to-A weights more than the S-to-B weights. If the network instead chooses action B, S-to-B connections have higher eligibility traces than S-to-A connections, but as no reward is given, no significant weight change occur. Therefore, the complexity of the problem is not combinatorial. Nevertheless, the present experiment shows that the modulated eligibility traces with the RCHP rule have the capability of determining very selectively which pathways are responsible for a reward and which pathways are not, even when the possible combinatorial choices of those random pathways are large in number.

4.4 Shift of modulatory response to earlier predicting stimuli

This section shows that the proposed neural model reproduces the shift of modulatory response from the initial unconditioned stimulus to earlier and predicting stimuli.

An unconditioned stimulus (US) is a stimulus that triggers an innate reaction. For example, in Pavlov's experiment food triggers salivation in the dog. A conditioned stimulus (CS) is a cue that precedes the US, as the ring of a bell shortly before the food is given. A remarkable feature of classical conditioning is that the response to the reward (i.e. salivation in the dog) shifts to the predicting stimuli (i.e. the ring of a bell), once the subject has learnt the association. The experiment in this section reproduces in simulation the shift of a modulatory response to two predicting and preceding conditioned stimuli as in Izhikevich (2007, Problem 4, Page 2448).

One hundred excitatory neurons are randomly selected in the network (group US) to encode an unconditioned stimulus. Another 100 random excitatory neurons form a group $\mathrm{MOD_p}$ (named VTA_P in Izhikevich (2007) where the subscription P stands for *projecting*) whose activity regulates the modulation level of the network, i.e. the activity of neurons $\mathrm{MOD_p}$ determines the intensity of d in Eq. 6. The weights from US to $\mathrm{MOD_p}$ are set to the maximum value to mimic the initial modulatory response to the US. This means that when a US is delivered, the US group, which connects strongly to the $\mathrm{MOD_p}$ group, causes high activity in the $\mathrm{MOD_p}$ group and consequently modulatory activity spreads throughout the network. Two other groups, each with 100 excitatory neurons (groups $\mathrm{CS1}$ and $\mathrm{CS2}$) are randomly chosen to encode two conditioned stimuli CS1 and CS2. Fig. 7a represents graphically the four groups of neurons.

Every 10 to 30 s, the network receives CS1 followed by US with a delay of 1 ± 0.3 s. After 100 trials, CS2 also occurs 1 ± 0.3 s before CS1. At the start of the simulation, the network shows a peak in the modulatory activity when US is received (Fig. 7b); this is due to the prewiring that set the connections from US to $\mathrm{MOD_p}$ to the maximum value. Stimulus CS1 does not elicit a modulatory response during the first trials. However, after only a few pairing episodes, CS1 also starts triggering modulatory activity. The modulatory peak progressively shifts from the US to the CS1 until it is stronger when a CS1 is delivered (trial 100, Fig. 7c). From trial 101 on, CS2 is also delivered 1 ± 0.3 s before CS1. Again the peak of modulatory activity shifts progressively from CS1 to the predicting CS2: at the end of the simulation, the modulatory activity is strongest when CS2 is delivered, the earliest predictor of US (Fig. 7d). This effect, claimed in Izhikevich (2007) to be due to the millisecond firing scale, is nevertheless reproduced here with the rate-based model, proving the hypothesis that the rarity of correlations is the driving mechanism for this type of response.

As in the experiments discussed in the earlier sections, these particular learning dynamics derive from the rare eligibility traces generated by the RCHP rule. In fact, each time the CS1 is delivered to the network, the connections from CS1 to the rest of the network produce eligibility traces, similarly to the classical conditioning experiment (Section 4.2). The following reward, generated by the US, then transforms those eligibility traces into long-term synaptic increase. This means that the connections from $\mathrm{CS1}$ to the rest of the network, and therefore also to US and $\mathrm{MOD_p}$ grow stronger. As a consequence, CS1 is capable of triggering modulatory activity alone, before the US is delivered. The same mechanism applies when CS2 occurs before CS1. The modula-

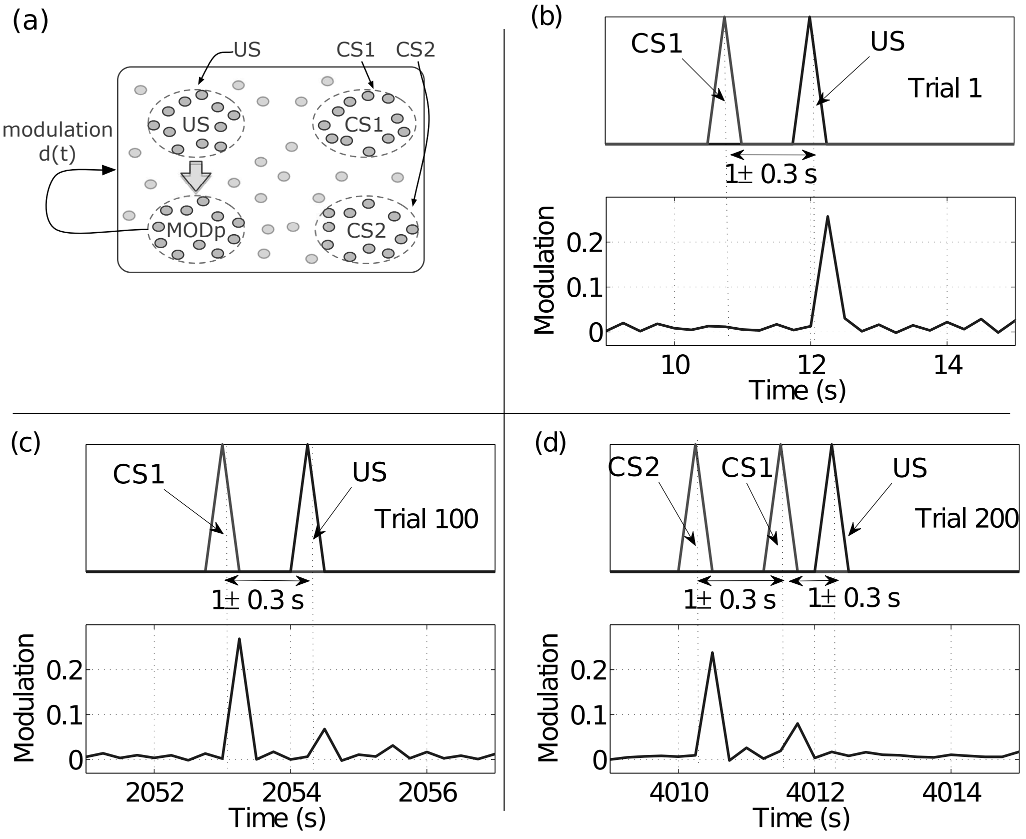

Figure 7: Rate-based implementation of the shift of modulation from US to reward-predicting stimuli in classical conditioning. (a) Graphical representation of the four groups of neurons that identify the unconditioned stimulus (US), the modulation-inducing neurons ($\mathrm{MOD_p}$) and the two groups of conditioned stimuli (CS1 and CS2). The group $\mathrm{MOD_p}$ is formed by excitatory neurons, but determines directly the level of modulation d of the whole network. This is done by assuming that MODp projects to a group of modulatory neurons (not modelled) which in turn modulates the network. (b) At the beginning of the simulation the US triggers a strong modulatory response. (c) After 100 trials during which the CS1 always predicts the US, the modulatory response has shifted to the predicting CS1. (d) At trial 200, after CS2 has been predicting CS1 and therefore US for 100 trials, the modulatory response has further shifted to the earliest predicting stimulus CS2. The simulation reproduces qualitatively the learning dynamics of the spiking neurons in Izhikevich (2007, Fig. 4, Page 2449).

tory peaks for the US drop when CS1 is learnt because the neurons in the group $\mathrm{MOD_p}$ become strongly driven by the neurons in $\mathrm{CS1}$. This means that each time CS1 is delivered, neurons in $\mathrm{MOD_p}$ respond strongly, thereby causing decorrelating traces in the synapses US-to-$\mathrm{MOD_p}$. In other words, the preceding CS1, that is an earlier predictor of the US, competes with the modulatory response to the US, which drops in intensity as in biological recordings (Pan et al., 2005).

These learning dynamics allow the randomly connected network of $1\,000$ neurons to simulate the stimulus-response shift of classical conditioning. Similarly to biological findings, the network is capable of predicting the delivery of a US and shifts the modulatory activity to the earliest predicting stimulus (Schultz et al., 1997; Schultz, 1998, 2006).

5 Implications of learning with rare correlations

The concept of rare correlations is exploited in this section to understand its role in the learning dynamics. In particular, four aspects are exposed: 1) the increase of the delay of rewards thanks to even rarer correlations; 2) the interpretation of short-term plasticity as a form of eligibility trace; 3) various decays of the eligibility traces and of the modulation and 4) the robustness of learning when modulation causes also an increase in excitatory activity.

5.1 The relationship between the rarity of correlations and the time-constant of traces: extending the time to reward

The time-constant in the range $[1, 2]$ s used in the previous experiments means that a trace decays to negligible values after 3-6 s according to the endogenous factors in Eqs.1 and 5. Rewards occurring after such an interval cannot strengthen the correct synapses. One solution may be that of increasing the time-constant of traces, which results in a slower decay and potentially allows for a longer delay of the rewards. However, if traces are generated at the rate of 1%/s, but last longer due to a slower decay, the result is that more traces are present in the network at any time, making more synapses eligible for update at the reward delivery. Additionally, if a reward has a large delay, the probability of decorrelations occurring between the action and the reward increases. Therefore, one hypothesis is that longer delays are possible only when traces have a slower decay and, at the same time, correlations are rarer. By making correlations ever rarer during a unit of time, fewer traces are generated or modified, thereby allowing for their extension in time.

To test this hypothesis, the experiment of reinforcing a synapse (section 4.1) is reproduced in three additional conditions to demonstrate the relationship between rarity of correlations and the decay of traces. The delay of rewards is made 15 times longer with an extended interval in $[1, 45]$ s. In a first experiment, all other settings are left unchanged. This test is intended to show that when the delay of rewards exceeds the duration of the traces, the learning breaks down. As predicted, Fig. 8a shows that the reward-triggering synapse σ is not reinforced: in the extended simulation time of 4 h,

127 reward episodes occurred, causing the synapse σ to grow only marginally and inconsistently. In fact, when a reward is delivered, the trace of the synapse σ has dropped to negligible values due to the long delay. The histogram shows that other synapses grow to larger values than σ, clearly indicating that the learning fails.

In a second simulation, the time-constant of the traces is also increased 15 times (i.e. to 30 s). Fig. 8b shows that the reward-triggering synapse σ is reinforced, but drops frequently to lower values. The occasional decreases are due to decorrelations that occur during the long delay between the action and the reward. In other words, with a maximum delay of 45 s and a time-constant of traces of 30 s, 1% correlations per second are not rare enough. Therefore, a third simulation is run with a rarer correlation rate of 0.2%/s. Fig. 8c shows that the learning dynamics are restored and are very similar to those of the initial experiment in section 4.1. Therefore, by increasing the time-constant of traces and reducing at the same time the rarity of correlations, it is possible to solve the distal reward problem with longer reward delays.

The results in this section prove that the rare correlations in the RCHP operate in combination with the decay rate of traces to maintain the correct learning dynamics. A decrement in the correlation rate and an increase in the time-constant of the traces maintain the correct balance of traces to learn with an extended reward delay in the range of 1 to 45 s. This experiment establishes for the first time a criterion to understand how different delays can be accounted for. This new insight promises to be fundamental in the implementation of artificial learning networks that cope with highly variable delays between actions and reward.

This principle of dependence between rarity of traces and their decay time effectively makes a prediction on neural dynamics. If fact, if traces have decayed by the time a reward is delivered, the distal reward problem cannot be solved. If, on the other hand, traces are too many, the correct synapses cannot be identified, or, alternatively, too many candidate synapses may grow to excessive values. In Izhikevich (2007), the time-constant of traces is constant at 1 s and the average firing rate is 1 Hz. However, other regimes, in which delays are longer or correlations are more frequent or rarer, are also conceivable. According to the result in this section, biological networks could regulate this mechanism by lowering the rate of production of traces when longer traces are necessary, i.e. when longer delays occur. This hypothesis could explain how classical and operant animal conditioning problems (Staddon, 1983) with longer delays are solved. Interestingly, this prediction of the rates of creation and extinction of traces cannot yet be tested in biological networks because the precise nature of traces, chemical or electrical, has not been exactly established. The biological mechanisms that generate and maintain traces remain also speculative. Nevertheless, the relationship presented in this section is a theoretical principle that holds its generality regardless of the precise nature of traces. This fact is also supported by the experiment in the next section, in which this generality is further demonstrated by showing that the learning is preserved when traces are not implemented by a chemical concentration but rather by the strength of the synapse itself.

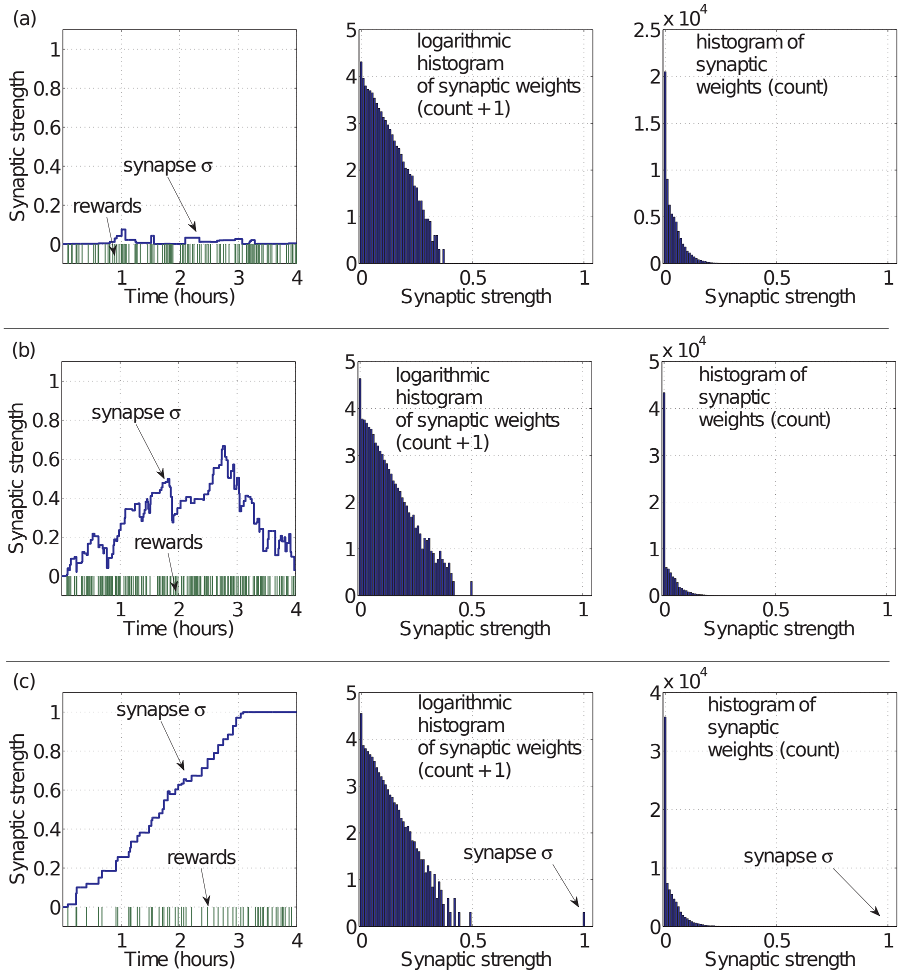

Figure 8: Relationship between the rarity of correlations and the time-constant of eligibility traces. In these simulations, the delay of rewards is extended in the interval $[1, 45]$ s, 15 times as long as before. (a) All parameters, including the correlation rate and the time-constant of traces are unchanged (respectively 1%/s and 2 s). The synapse σ is not significantly reinforced despite more than one hundred reward episodes (b) The time-constant of traces is extended to 30 s. The synapse σ is reinforced, but not consistently. (c) The rate of correlations is reduced to 0.2%/s. The potentiation of the synapse σ is correctly achieved even with the extended delays of rewards to 45 s.

5.2 The weight strength: a more plausible eligibility trace

Eligibility traces express a *tag* or an indication that a synapse is eligible for growth. However, it is not clear whether or how this is reflected in biological neural networks. The approach in Izhikevich (2007) is to model a synaptic-specific concentration of a chemical that enables weight increase only in the presence of modulation, but not when STDP occurs (Fig. 1). However, the biological evidence for chemical traces is at best incomplete (Magee and Johnston, 1997; Markram et al., 1997; Bailey et al., 2000); for example, the effect of the modulatory chemical acetylcholine is that of consolidating synaptic growth, but short-term potentiation due to STDP is observed even without modulation (see Bailey et al. (2000) for a review and hypotheses on heterosynaptic plasticity). In Bailey et al. (2000), a number of reviewed studies show that STDP can lead to short-term growth, without modulation, followed by a fast weight decay. Modulatory activity has the capability of consolidating such growth to last in the long term. The experiment in this section models this phenomenon within the current neural network.

In this section, the model is modified such that a correlating event, detected by the RCHP rule, increases the weight. Eligibility traces are eliminated. However, the increment given by the RCHP rule is added to the weight with a fast-decaying nature. If no reward occurs within a few seconds, the weight returns to its original value. If a reward occurs, the fast-decaying component is consolidated into a slow-decaying or permanent weight. Therefore, the weight strength, in this section indicated with W to distinguish it from the notation of the previous sections, is expressed as the sum of two components: a fast-decaying short-term weight w_{st} and a long-term weight w_{lt}. The short-term, long-term and overall components of the weight are expressed by

$$\begin{aligned} \dot{w}_{st} &= -w_{st}/\tau_c + \text{RCHP} && (9)\\ \dot{w}_{lt} &= d \cdot w_{st} && (10)\\ W &= w_{st} + w_{lt} && (11) \end{aligned}$$

where τ_c is the same time-constant used earlier for eligibility traces. The overall weight W is therefore the sum of a base-level value given by w_{lt} that changes only in the presence of modulation, and a more fluctuating term w_{st} that changes with RCHP alone and decays quickly. The differential Eq. 9 is integrated as previously done with Eqs. 1 and 5.

The first experiment of reinforcing one synapse (Section 4.1) is run again with the new weight update rule of Eqs. 9-11 and no eligibility traces. The synapse σ grows to reach the maximum value throughout the simulation (Fig. 9). The weight also shows high fluctuations due to the short-term component. At the end of the simulation, the overall distribution of the long-term components of the weights (middle plot in Fig. 9) shows that the learning is strikingly similar to the experiment with the eligibility traces. The histogram of the weights, instead, (plot to the right in Fig. 9) displays larger weight values, which are the sum of the long- and short-term components. The weight dynamics are visualised graphically with a snapshot of the connection σ during a brief interval (Fig. 10). The weight strength is shown in its long- and short-term components.

The fundamental implication of this last experiment is that the weights themselves can serve as eligibility traces. Further studies are needed to assess whether separate el-

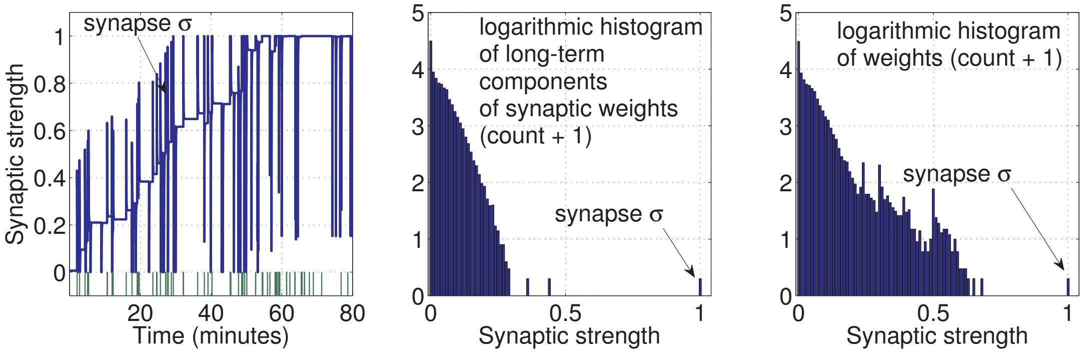

Figure 9: Instrumental conditioning of one synapse in which the short-term weight component represents the eligibility trace. The plot on the left shows the weight σ that increases during the simulation and reaches the maximum allowed value while displaying at times large variations. The middle plot is the histogram of the sole long-term component of all weights. The plot on the right is the histogram of weights, i.e. the sum of long- and short-term components.

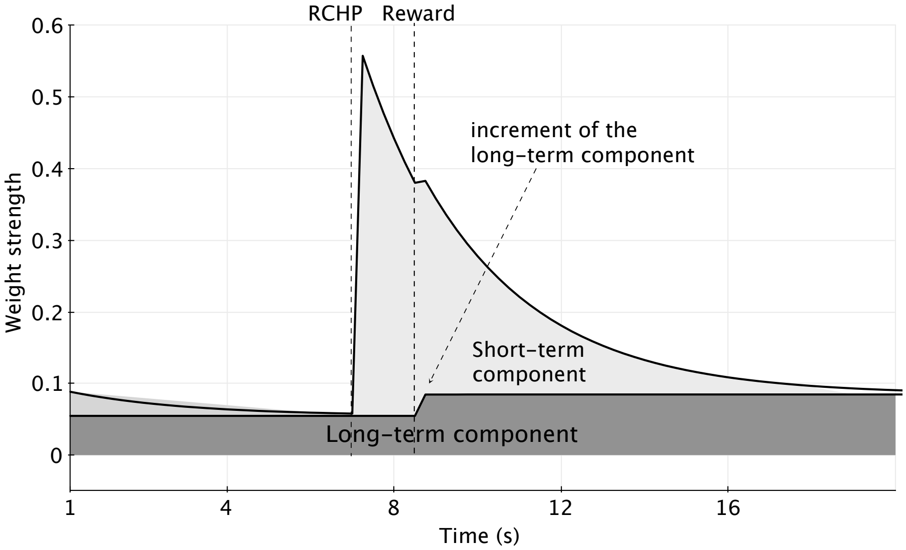

Figure 10: Snapshot of a brief simulation interval on the weight σ graphically decomposed into its long- and short-term components. A correlation event, detected by the RCHP rule, increases the short-term component (light grey area). The short-term component decays quickly and no long-term changes take place unless a reward is delivered. The subsequent reward modifies the long-term component (dark grey area) by consolidating the short-term component in proportion to the intensity of the modulation (see Eq. 10). Therefore, the RCHP alone changes the short-term component, but cannot alter the long-term component. Reward alone also cannot change the weight strength. The combination of RCHP and following reward repeated several times result in an overall increase of the weight strength.

igibility traces are advantageous in certain situations. However, a simple consideration is that using the weight itself as an eligibility trace introduces more exploratory potential. In fact, with eligibility traces, as in Izhikevich (2007), a weight does not increase, and therefore cannot grow to a large value, unless a reward is delivered. In contrast, in this last experiment, a weight that experiences repeated correlations may grow to a large, although short-lived, value. Any weight in the network may grow temporarily to a large value due to correlating, but random, activity. Therefore, if the action that triggers a reward requires a large weight, for example because a strong output response is required, this latest approach may succeed. In short, this second approach of using weights as eligibility traces, which is proven effective, might be beneficial in exploring a wider range of neural states, possibly resulting also in a wider exploration of the action space.

The fact that a weight can be temporarily increased by repeated stimulations, but returns to its original strength shortly after, is a common finding in biological measurements (Bailey et al., 2000). Therefore, this version of the model, in which the weight has a short- and a long-term components, implicitly suggests a possible computational function of short- and long-term plasticity. The time scale of plasticity, i.e. short-term and long-term, is often related to the duration of time that the information is to be preserved in the network. The current model instead implements a mechanism in which short-term plasticity represents eligible information to be possibly transformed into long-term memory.

Finally, the fact that the learning dynamics are preserved when eligibility traces are represented by the short-term plasticity of synapses suggests that the principle of rarity is independent of the particular form that traces assume. For example, sustained firing and reverberating activity (Hebb, 1949; Histed et al., 2009; Pawlak et al., 2010) are alternative factors that have been suggested to encode eligibility traces. The reply of behavioural sequences has also been suggested to help consolidate reward-based learning (Foster and Wilson, 2006). The principle of rarity suggests that, even in the interesting case that eligibility traces are encoded by activity, the average number of eligible synapses is a critical factor, and according to the present study must be maintained at a low percentage to ensure stability and learning.

5.3 Modelling various decays of modulatory signals and traces

The modulatory signal is encoded in the current model as a single-step input signal representing the timing of reward. This signal multiplies the eligibility traces to determine the weight updates. Therefore, the timing and decay of both traces and modulatory signals are important factors in determining the weight changes. This section investigates in detail how the weight change is affected by various decays in both the trace and the modulatory signal.

In a first instance, three types of decay of the modulatory signal are modelled: 1) a biological plausible decay of 0.2 s, similar to that measured in Wighmann and Zimmerman (1990); Garris et al. (1994); Cass and Gerhardt (1995); 2) a longer decay of 1 s scaled in amplitude to preserve the total amount of signal per reward episode and 3) a decay of 1 s without scaling, representing case (1) with a slower reuptake. These three types of decay are tested in a simple simulation of one correlating episode at one

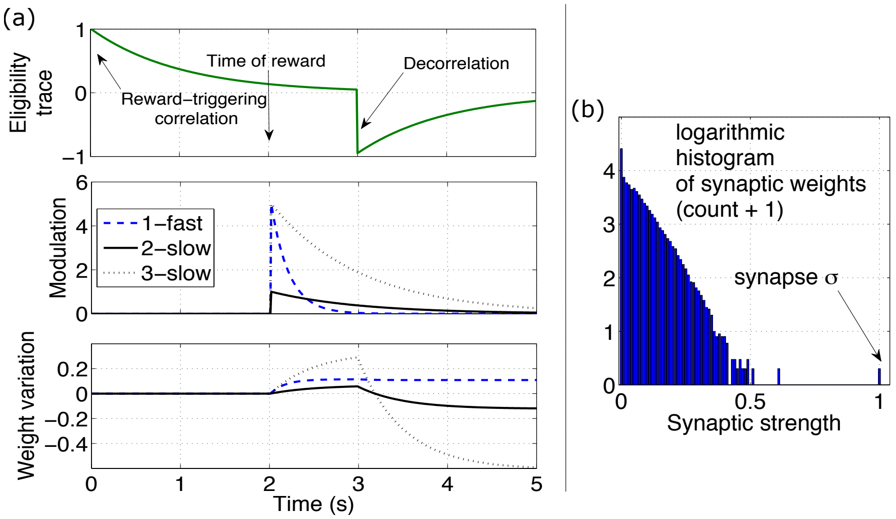

Figure 11: Effect of various decays of modulation on the weight change. (a) An exponentially decaying eligibility trace is multiplied by three different types of modulatory decays. The modulations with long decays cause the initial weight increase to be reduced by the further decorrelation episode at 3 s. (b) Weight distribution after 1.5 h of simulated time with a modulatory signal with a time-constant of 1 s.

synapse. The correlation creates a trace followed by a reward after 2 s. After one further second, a decorrelation occurs. The initial correlation and the following reward represent the event that reinforces the weight. The following decorrelation represents a random noise-induced event. The purpose is to test how the three types of decay affect the weight update.

Fig. 11a shows the eligibility trace in the first row, the three types of modulatory signals in the second row, and the products representing the weight change in the third row. The final weight change is positive with the fast modulatory signal (case 1), while it is affected negatively by the decorrelation episodes when the modulation has a slow decay (cases 2 and 3). The reason is that the fast modulatory signal causes a quick conversion of the trace into a weight change. When such a process takes longer, due to a slower modulatory signal, unrelated neural activity may disrupt the correct weight change. It is nevertheless interesting to test to which extent a long modulatory signal with a slow decay disrupts the learning in the experiment of reinforcing one synapse. Fig. 11b shows the histogram of the final weights after 1.5 h of simulated time with a modulatory signal with a 1 s time-constant (case 2). The synapse σ is the only one to reach saturation, proving the robustness of the algorithm in such conditions.

It is worth noting that a modulatory signal with a time-constant of 1 s has a slow decay with respect to traces that also decay at a similar rate. In the case of longer-lasting traces, as in the case of the experiment in section 5.1 in which the time-constant is 30 s, a modulatory signal that decays with a 1 s time-constant is relatively fast. Therefore, from a computational perspective, it can be inferred that slow modulatory signals are efficient on even slower neural dynamics, e.g. eligibility traces that are persistent on a longer time scale.

The traces allow for the reconstruction of the cause-effect relationship between cor-

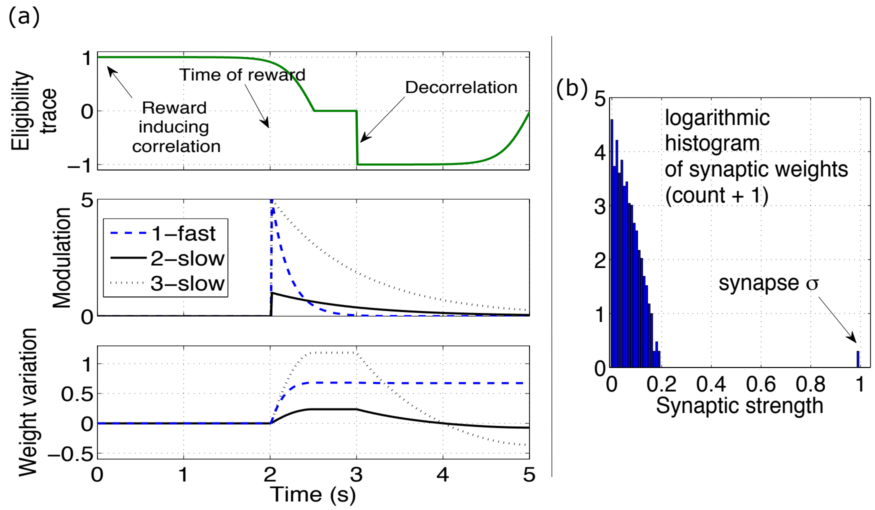

Figure 12: Effect of various decays of the eligibility traces on the weight change. (a) An eligibility trace with a decay regulated by the hyperbolic tangent function is multiplied by three different modulatory decays. In this case, as for the exponential traces, the weight increase is maximised and preserved with the fast decaying modulatory signal. (b) Weight distribution after 1.5 h of simulated time with 1-step modulation and traces matching the probability distribution of reward delivery.

relations and rewards. Therefore, the dynamics of the trace are also fundamental in the weight change. Fig. 12a shows a differently shaped trace which decays as the hyperbolic tangent. This type of trace, when multiplied by the modulation, provides a greater weight change with the fast modulatory signal, i.e. it appears more effective. Interestingly, all previous experiments used an exponentially decaying trace, but the distribution of rewards (in the experiment of section 4.1) is uniform in the interval $[1, 3]$ s. The implication is that the trace is highest when the reward does not occur, i.e. between 0 and 1 s after the correlation. Therefore, one hypothesis is that an efficient procedure to solve the distal reward problem matches the decay of the traces with the probability distribution of rewards. In this way, the traces will be maximum during the expected time-window of reward delivery. Accordingly, in a new test, the shape of the trace is matched to the probabilistic distribution of the reward delays. I.e., the traces are a constant positive value in the interval $[1, 3]$ s after a correlation occurs, and zero otherwise. Negative traces are left unchanged, i.e. they decay exponentially. Fig. 12b shows a great separation between the reinforced synapse σ and the other synapses in the network. The low values of the synapses also imply that the unwanted variation of other synapses that are not involved in the reward-triggering process is greatly reduced.

The analysis in this section indicates that short-lived modulatory episodes are more suited in the integration of traces than slow-decaying modulation. The possible biological implication is that slow modulatory signals may act on even slower neural dynamics, such as long-lasting eligibility traces. The simulation with a trace matching the distribution of reward delays also proved that a correspondence between the distribution of the reward delays and the shape of traces is determinant in improving the efficacy in the solution of the distal reward problem. In this case, the experiment predicts that efficient biological neural dynamics for the solution of the distal reward problem must be matching the decay of traces with the probability distribution of future rewards.

5.4 Robustness to divergence of neuromodulators

The computational role of neuromodulation is that of representing a different type of signal from neural activation. For this reason, the modulatory signal that modulates plasticity is sometimes called a third factor (Porr and Wörgötter, 2007), after the presynaptic and postsynaptic factors, that stabilises Hebbian plasticity. If modulation causes excitation, the two types of signals are not separated anymore, and the learning can be expected to break down. However, one may ask if this separation of signal has to be absolute, or whether a certain level of dependence between neural activation and modulation can occur without disrupting the reward learning.

The biological ground for this question lies in the concept of divergence of neurotransmitters. Generally, each type of neurotransmitter binds to a specific receptor. Exceptions to this rule result in a property called *divergence*, which is caused by one neurotransmitter binding to more types of receptors (Bear et al., 2005). This means that a modulatory neurotransmitter may affect simultaneously different properties of the synaptic junction, e.g. plasticity, efficacy, or other neural states. Additionally, the Dale's principle (Dale, 1935; Strata and Harvey, 1999), according to which each type of neuron releases only one type of neurotransmitter, has exceptions across the variety of neuron types in the brain (Bear et al., 2005). This condition is modelled in the present section by introducing two variations in the model. In the first, a modulatory peak causes all $1\,000$ neurons in the network, i.e. all neurons undergoing neuromodulation, to receive a large excitatory input of 5. In a second variation, a modulatory peak doubles the activation factor γ of Eq. 8, i.e. it increases the efficacy of all neurons in the network during the time step at which modulation is delivered. Therefore, in both extensions, the modulation has a combined effect on plasticity and neural activity.

The first simulation produces the learning dynamics in Fig. 13a and Fig. 13b. The plots indicate that the increase of neural activity by modulation does not affect significantly the learning produced by rare correlations. Interestingly, the increased neural activity caused by modulation leads also to an increase of the rate of rare correlations at the moment of reward delivery (Fig. 13b). As opposed to Fig. 4, the percentage of rare correlations exceed considerably the target range of $[0.5$,$1.5]$%/s. However, as the peaking episodes are occasional, the learning is preserved. The learning is also preserved because the large amount of traces generated by the modulatory-induced activity have a later onset than the modulation. In other words, the modulatory signal increases the overall excitatory activity of the network, but this excitatory wave follows in time the modulatory peak.

Fig. 13c shows the learning dynamics when modulation increases the gain in the network. The learning does not appear to be affected by the modulatory effect on the neural gain. However, it is important to note that the gain is only temporarily increased by modulation. A permanently high gain may affect the weight stability by triggering self sustained activity.

Finally, it is important to note that the proposed model prescribes that correlations are rare on average, i.e. during an interval of time, and across the whole network. Therefore, the network activity may display unusually high levels of correlations in particular conditions, e.g. when a stimulus is delivered, or at particular locations, e.g. over a reward-triggering pathway. In short, temporary variations of the activity and gain are effectively handled as disturbing stimuli and are proven in this section not to affect the correct learning dynamics. The ability of the network to reinforce only the pathways that are causally related to a subsequent reward is further demonstrated.

Conclusion

The current study identifies the principle of rare correlations as a pivotal element in neural learning to solve the distal reward problem. Rare correlations are a means to create few eligibility traces, which, over many reward episodes, are functional in isolating the reward-triggering pathways. Rare correlations are detected in the model of this study by means of a new formulation of the Hebbian rule, named Rarely Correlating Hebbian Plasticity (RCHP). The new rule, in combination with neuromodulation, is shown to solve the distal reward problem in a variety of experimental scenarios. The learning is achieved with a rate-based model across a large range of sampling steps, thereby rejecting a previous hypothesis that the precise spike-timing of spiking neurons was required to solve this problem (Izhikevich, 2007).

The application of the principle of rare correlations allows neural models to cope with temporal gaps between actions and rewards and to associate rewards with previous actions and cues even in the presence of intervening, disturbing stimuli. Classical, instrumental conditioning and the shift of the modulatory response to conditioned stimuli is demonstrated in simulation. The instrumental conditioning scenarios show the ability of the network to learn from its own actions even when rewards occur with delays of variable duration. The experiments in classical conditioning demonstrate the predictive ability of the network, which can associate stimuli that are causally related even when the intervening time is uncertain, and disturbing stimuli occur in between.

The analysis shows that a balance between the rarity of correlations and the duration of traces is essential for the network's stability and correct learning. By decreasing further the probability of correlations, traces with longer decays can be used to account for longer delays between actions and rewards. For the first time, a criterion that explains how to account for long delays of the reward is proposed. While the precise nature of traces in the brain is not fully established, the proposed model predicts that a correct balance between the rate of generation and the rate of expiration of the traces must be maintained.

The independence of the principle from the exact nature and dynamics of the traces is shown with the use of short-term weight updates in lieu of traces. In fact, when the synapse-specific chemical, which represents the eligibility trace, is replaced by short-term plasticity in the synaptic weights themselves, the learning is preserved even though the weights display a higher degree of variation. This finding suggests a new interpretation of short- and long-term synaptic plasticity in biological networks. Rather than representing memory in the short or long term, could short-term plasticity represent a way of exploring network functions and making synapses eligible for long-term storage? Further research in biological neural networks is needed to resolve this question.

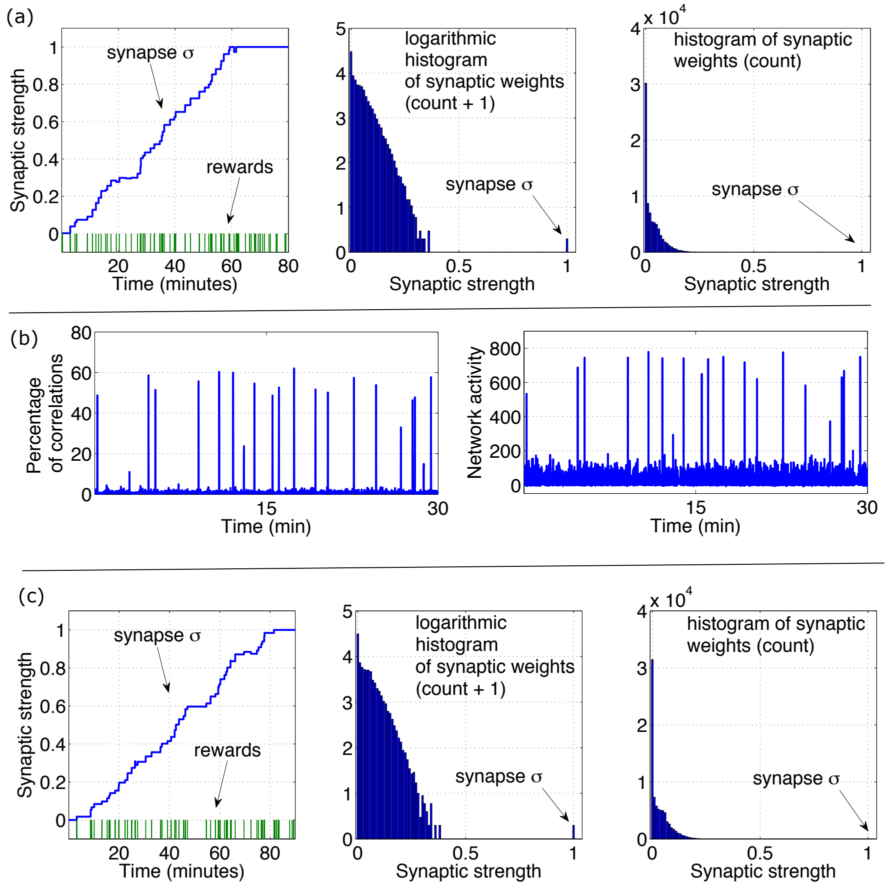

Figure 13: Learning with variations of the model that cause modulation to affect neural activity. (a) Weight growth of σ and histograms of the final weight configurations in the experiment in which the modulation causes an increment of the perceived input by 5 for all neurons in the network. (b) Percentage of rare correlations and activity of the whole network during the experiment in (a). Each reward-delivery increases the activation of all neurons in the network, therefore causing a temporary peak in the number of correlations. (c) Weight growth of σ and histograms of the final weight configurations in the experiment in which the gain γ of all neurons in the network is doubled when a modulatory signal occurs.

The final experiments show that the efficiency in solving the distal reward problem is increased when the decay of traces matches the probability distribution of the delayed rewards. The robustness of the principle is demonstrated across all experimental scenarios.

In conclusion, the principle of rare correlations identified in this study, and the corresponding novel Hebbian rule, bridge long-term memory consolidation with faster neural processes, allowing a network to perform classical and instrumental conditioning with asynchronous events. This finding encourages a wider interpretation of biological synaptic plasticity. It also invites novel experimentations of neural learning in simulated or neuro-robotic scenarios, particularly real-time conditions where the outcome of actions is generally known with variable delays. The concept of rare correlations casts a new light on the computational possibilities of traditional Hebbian plasticity and reveals unforeseen roles in reward and associative learning.

Acknowledgments

Special thanks to John A. Bullinaria, Peter Duerr, Billy Land, Patty Land, Andre Lemme, Ben H. Jones, Felix Reinhart, Kenneth O. Stanley and Paul Tonelli for valuable comments on earlier drafts of this paper. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013, Challenge 2 Cognitive Systems, Interaction, Robotics under grant agreement No 248311 - AMARSi.

Appendix

5.5 Implementation details

A summary of the settings for each experiment is provided in the tables of this section. Table 1 summarises the simulation parameters that are common to all experiments.

In particular, the connection probability (0.1) can be implemented as: (a) any two neurons have 0.1 probability of being connected; (b) any neuron has 100 input connections from random neurons; (c) any neuron has 100 output connections to random neurons. The cases (b) and (c) with exactly 100 inputs or outputs produce more regularly connected networks. When any two neurons are connected with probability 0.1 (case (a)), the network is more irregularly connected with some neurons having the number of inputs as low as 70 or as high as 130. Simulations revealed that, in this latter case, the results were qualitatively similar to the case with a regular connectivity, but they showed more variation in the time of the learning process. The outcome of the simulations with different random initialisations (provided as support material) were produced with initialisation (b). Additionally, neurons do not connect to themselves, and two neurons do not have multiple connections.

The sampling time step is varied in the experiments in Section 4.1 from 10 to 1 000 ms to show that the precise time of the computation is not a crucial element in the learning dynamics that solve the distal reward problem. The maximum modulatory value (0.12) determines the amount of trace that is converted in weight change according to Eq. 6. Higher or lower values determine a faster or slower learning rate.

Excitatory neurons	800
Inhibitory neurons	200
Connection probability	0.1
Weight range	$[0, 1]$
Inhibitory weights	Fixed in $[0, 1]$
Excitatory weights	Plastic
Noise on neural transmission	Uniform $[-0.15, 0.15]$
Target rate of rare correlations	1%
Sampling time step	$[10, 1000]$ ms
Time-constant of eligibility traces (τ_c)(*)	$1 - 30$ s
Neural gain γ of Eq. 5	0.2
Maximum modulatory value	0.12

Table 1: Summary of parameters across all simulations. (*) The time-constant of traces is varied across the experiments, see tables below.

Simulation time	5 400 s
Delay of reward	[1, 3] s
Time constant of eligibility traces (τ_c)	2 s
Sampling time step	$[10, 1000]$ ms

Table 2: Specific parameters for the simulation of Section 4.1 “Reinforcing a synapse”.

The following Tables 2, 3, 4 and 5 refer respectively to the experiments in Sections 4.1, 4.2, 4.3, 4.4. Section 5.1 uses the parameters summarised in Table 6. Section 5.2 has identical settings are Section 4.1 “Reinforcing a synapse” in which the weight update is given by Eq. 11.

The algorithms and simulations presented in this study are implemented and run with Matlab scripts provided as a support material. The scripts can be downloaded at http://andrea.soltoggio.net/RCHP.

Simulation time	5 400 s
Delay of reward	$[0, 1]$ s
Time constant of eligibility traces (τ_c)	1 s
Sampling time step	25 ms
Number of stimuli	100
Stimulus represented by	50 random excitatory neurons
Stimulus strength (added to u)	$+20$
Inter-stimuli time	$[100, 300]$ ms

Table 3: Specific parameters for the simulation of Section 4.2 “Classical (Pavlovian) Conditioning”.

Simulation time	1 000 s (100 stimuli)
Delay of reward	$[0, 1]$ s
Time constant of eligibility traces (τ_c)	1 s
Sampling time step	100 ms
Stimulus represented by	50 random excitatory neurons
Stimulus strength (added to u)	$+20$
Inter-stimuli time	10 s

Table 4: Specific parameters for the simulation of Section 4.3 “Instrumental Conditioning”.

Simulation time	3 000 s
Interval between CS and US	$[0.7, 1.3]$ s
Time constant of eligibility traces (τ_c)	1 s
Sampling time step	100 ms
Stimulus represented by	100 random excitatory neurons
Stimulus strength (added to u)	$+20$
Inter-stimuli time	$[10, 30]$ s

Table 5: Specific parameters for the simulation of Section 4.4 “Shift of modulatory response to earlier predicting stimuli”.

Experiment	Time-constant of traces	Rare correlations	Reward delay
In Fig. 8a	2s	1%/s	[1-45] s
In Fig. 8b	30s	1%/s	[1-45] s
In Fig. 8c	30s	0.2%/s	[1-45] s

Table 6: Specific parameters for the simulation of Section 5.1 “The relationship between the rarity of correlations and the time-constant of traces: extending the time to reward”.

References

Abbott, L. F. (1990). Modulation of Function and Gated Learning in a Network Memory. *Proceedings of the National Academy of Science of the United States of America*, 87(23):9241–9245.

Alexander, W. H. and Sporns, O. (2002). An Embodied Model of Learning, Plasticity, and Reward. *Adaptive Behavior*, 10:143.

Arbuthnott, G. W. and Wickens, J. (2007). Space, time and dopamine. *Trends in Neurosciences*, 30(2):62–69.

Bailey, C. H., Giustetto, M., Huang, Y.-Y., Hawkins, R. D., and Kandel, E. R. (2000). Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? *Nature Reviews Neuroscience*, 1(1):11–20.

Barco, A., Lopez de Armentia, M., and Alarcon, J. M. (2008). Synapse-specific stabilization of plasticity processes: The synaptic tagging and capture hypothesis revisited 10 years later. *Neuroscience and Biobehavioral Reviews*, 32:831–851.

Bear, M. F., Connors, B. W., and Paradiso, M. A. (2005). *Neuroscience: Exploring the Brain*. Baltimore, MD.; London : Williams & Wilkins.

Bi, G.-q. and Poo, M.-m. (1998). Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *The Journal of Neuroscience*, 18(24):10464–10472.

Bi, G.-q. and Poo, M.-m. (2001). Synaptic Modification by Correlated Activity: Hebb's Postulate Revisited. *Annual Review of Neuroscience*, 24:139–166.

Cass, W. A. and Gerhardt, G. A. (1995). In Vivo Assessment of Dopamine Uptake in Rat Medial Prefrontal Cortex: Comparison with Dorsal Striatum and Nucleus Accumbens. *Journal of Neurochemistry*, 65:201–207.

Cooper, S. J. (2005). Donald O. Hebb's synapse and learning rule: a history and commentary. *Neuroscience and Biobehavioral Reviews*, 28(8):851–874.

Dale, H. H. (1935). Pharmacology and nerve-endings. *Proc. R. Soc. Med.*, 28:319–332.

Deco, G. and Rolls, E. T. (2005). Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex*, 15:15–30.

Farries, M. A. and Fairhall, A. L. (2007). Reinforcement Learning With Modulated Spike Timing-Dependent Synaptic Plasticity. *Journal of Neurophysiology*, 98:3648–3665.

Fellous, J.-M. and Linster, C. (1998). Computational Models of Neuromodulation. *Neural Computation*, 10:771–805.

Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19:1468–1502.

Foster, D. J. and Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(30):683–680.

Frey, U. and Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(533-536).

Garris, P., Ciolkowski, E., Pastore, P., and Wighmann, R. (1994). Efflux of dopamine from the synaptic cleft in the nucleus accumbens of the rat brain. *The Journal of Neuroscience*, 14(10):6084–6093.

Gerstner, W. and Kistler, M. W. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics*, 87:404–415.

Harris-Warrick, R. M. and Marder, E. (1991). Modulation of neural networks for behavior. *Annual Review of Neuroscience*, 14:39–57.

Hasselmo, M. E. (1995). Neuromodulation and cortical function: modeling the physiological basis of behavior. *Behavioural Brain Research*, 67:1–27.

Hasselmo, M. E. (2005). Expecting the unexpected: Modeling of neuromodulation. *Neuron*, 46(4):426–528.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York.

Histed, M. H., Pasupathy, A., and Miller, E. K. (2009). Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. *Neuron*, 63(2):146–148.

Hull, C. L. (1943). *Principles of behavior*. New-Your: Appleton Century.

Izhikevich, E. M. (2006). Polychonization: Computation with spikes. *Neural Computation*, 18(2):245–282.

Izhikevich, E. M. (2007). Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling. *Cerebral Cortex*, 17:2443–2452.

Kandel, E. R. and Tauc, L. (1965). Heterosynaptic facilitation in neurones of the abdominal ganglion of Aplysia depilans. *The Journal of Physiology*, 181:1–27.

Legenstein, R., Pecevski, D., and Maass, W. (2008). A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS Comput Biol*, 4(10):e1000180.

Magee, J. C. and Johnston, D. (1997). A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons. *Science*, 275.

Marder, E. (1996). Neural modulation: Following your own rhythm. *Current Biology*, 6(2):119–121.

Marder, E. and Thirumalai, V. (2002). Cellular, synaptic and network effects of neuromodulation. *Neural Networks*, 15:479–493.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. *Science*, 275:213–215.

Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, 377:725–728.

O'Doherty, J. P., Kringelbach, M. L., Rolls, E. T., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, 4(1):95–102.

Pan, W.-X., Schmidt, R., Wickens, J. R., and Hyland, B. I. (2005). Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network. *The Journal of Neuroscience*, 25(26).

Päpper, M., Kempter, R., and Leibold, C. (2011). Synaptic tagging, evaluation of memories, and the distal reward problem. *Learning & Memory*, 18:58–70.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford : Oxford University Press.

Pawlak, V., Wickens, J. R., Kirkwood, A., and Kerr, J. N. (2010). Timing is not Everything: Neuromodulation Opens the STDP Gate. *Frontiers in Synaptic Neuroscience*, 2.

Pfeiffer, M., Nessler, B., Douglas, R. J., and Maass, W. (2010). Reward-modulated Hebbian Learning of Decision Making. *Neural Computation*, 22:1–46.

Porr, B. and Wörgötter, F. (2007). Learning with Relevance: Using a third factor to stabilize Hebbian learning. *Neural Computation*, 19(10):2694–2719.

Potjans, W., Diesmann, M., and Morrison, A. (2011). An Imperfect Dopaminergic Error Signal Can Drive Temporal-Difference Learning. *PLoS Computational Biology*, 7(5):1–20.

Potjans, W., Morrison, A., and Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. *Neural Computation*, 21(2):301–339.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, 58:322–339.

Redondo, R. L. and Morris, R. G. M. (2011). Making memories last: the synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, 12:17–30.

Rolls, E. T. (2009). *Handbook of Reward and Decision Making*, chapter From reward value to decision-making: neuronal and computational principles. Academic Press: New York.

Rolls, E. T., McCabe, C., and Redoute, J. (2008). Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. *Cerebral Cortex*, 18:652–663.

Sarkisov, D. V. and Wang, S. S. H. (2008). Order-Dependent Coincidence Detection in Cerebellar Purkinje Neurons at the Inositol Ttrisphosphate Receptor. *The Journal of Neuroscience*, 28(1):133–142.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80:1–27.

Schultz, W. (2002). Getting Formal with Dopamine and Rerward. *Neuron*, 36:241–263.

Schultz, W. (2006). Behavioural Theories and the Neurophysiology of Reward. *Annual Review of Psychology*, 57:87–115.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli during Successive Steps of Learning a Delayed Response Task. *The Journal of Neuroscience*, 13:900–913.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate for Prediction and Reward. *Science*, 275:1593–1598.

Skinner, B. F. (1953). *Science and Human Behavior*. New York, MacMillan.

Soltoggio, A., Bullinaria, J. A., Mattiussi, C., Dürr, P., and Floreano, D. (2008). Evolutionary Advantages of Neuromodulated Plasticity in Dynamic, Reward-based Scenarios. In *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*. MIT Press.

Soltoggio, A. and Stanley, K. O. (2012). From Modulated Hebbian Plasticity to Simple Behavior Learning through Noise and Weight Saturation. *Neural Networks*, 34:28–41.

Soula, H., Alwan, A., and Beslon, G. (2005). Learning at the edge of chaos : Temporal coupling of spiking neurons controller for autonomous robotic. In *Proceedings of the AAAI Spring Symposia on Developmental Robotics*.

Staddon, J. E. R. (1983). *Adaptive Behaviour and Learning*. Cambridge University Press.

Strata, P. and Harvey, R. (1999). Dale's principle. *Brain Research Bulletin*, 59:349–350.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA.

Thorndike, E. L. (1911). *Animal Intelligence*. Macmillan.

Timberlake, W. and Lucas, G. A. (1985). The basis of superstitious behavior: chance contingency, stimulus substitution, or appetitive behavior? *Journal of Experimental Analysis of Behaviour*, 44(3):279–299.

Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., and Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: When policy gradient methods fail. *PLoS Computational Biology*, 5(12).

Wang, S. S. H., Denk, W., and Häusser, M. (2000). Coincidence detection in single dendritic spines mediated by calcium release. *Nature Neuroscience*, 3(12).

Wighmann, R. and Zimmerman, J. (1990). Control of dopamine extracellular concentration in rat striatum by impulse flow and uptake. *Brain Res Brain Res Rev*, 15(2):135–144.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5:1–12.