Short-term plasticity as cause-effect hypothesis testing in distal reward learning

Andrea Soltoggio

Abstract Asynchrony, overlaps and delays in sensory-motor signals introduce ambiguity as to which stimuli, actions, and rewards are causally related. Only the repetition of reward episodes helps distinguish true cause-effect relationships from coincidental occurrences. In the model proposed here, a novel plasticity rule employs short and long-term changes to evaluate hypotheses on cause-effect relationships. Transient weights represent hypotheses that are consolidated in long-term memory only when they consistently predict or cause future rewards. The main objective of the model is to preserve existing network topologies when learning with ambiguous information flows. Learning is also improved by biasing the exploration of the stimulus-response space towards actions that in the past occurred before rewards. The model indicates under which conditions beliefs can be consolidated in long-term memory, it suggests a solution to the plasticity-stability dilemma, and proposes an interpretation of the role of short-term plasticity.

Keywords short-term plasticity · transient weights · distal reward · operant learning · plasticity vs stability · memory consolidation

1 Introduction

Living organisms endowed with a neural system constantly receive sensory information and perform actions. Occasionally, actions lead to rewards or punishments in the near future, e.g. tasting food after following a scent (Staubli et al, 1987). The exploration of the stimulus-action patterns, and the exploitation of those patterns that lead to rewards, was observed in animal behavior and named *operant conditioning* (Thorndike, 1911; Skinner, 1953). Mathematical abstractions of operant conditioning are formalized in algorithms that maximize a reward function in the field of reinforcement learning (Sutton and Barto, 1998). The maximization of reward functions was also implemented in a variety of neural network models (Lin, 1993; Pennartz, 1997; Schultz et al, 1997; Bosman et al, 2004; Xie and Seung, 2004; Florian, 2007; Farries and Fairhall, 2007; Baras and Meir, 2007; Legenstein et al, 2010; Frémaux et al, 2010; Friedrich et al, 2010), and is inspired and justified by solid biological evidence on the role of neuromodulation in associative and reward learning (Wise and Rompre, 1989; Schultz et al, 1993; Swartzentruber, 1995; Pennartz, 1996; Schultz, 1998; Nitz et al, 2007; Berridge, 2007; Redgrave et al, 2008). The utility of modulatory dynamics in models of reward learning and behavior is also validated by closed-loop robotic neural controllers (Ziemke and Thieme, 2002; Sporns and Alexander, 2002; Alexander and Sporns, 2002; Sporns and Alexander, 2003; Soltoggio et al, 2008; Cox and Krichmar, 2009).

Neural models encounter difficulties when delays occur between perception, actions, and rewards. A first issue is that a neural network needs a memory, or a trace, of previous events in order to associate them to later rewards. But a second even trickier problem lies in the environment: if there is a continuous flow of stimuli and actions, unrelated stimuli and actions intervene between causes and rewards. The environment is thus ambiguous as to which stimulus-action pairs lead to a later reward. Concomitant stimuli and actions also introduce ambiguity. In other words, any learning algorithm faces a condition in which one single reward episode does not suffice to understand which of the many preced-

Andrea Soltoggio
Computer Science Department
Loughborough University, LE11 3TU, Loughborough, UK.
E-mail: a.soltoggio@lboro.ac.uk

ing stimuli and actions are responsible for the delivery of the reward. This problem was called the *distal reward problem* (Hull, 1943), or *credit assignment problem* (Sutton, 1984; Sutton and Barto, 1998). Credit assignment is a general machine learning problem. Neural models that solve it may help clarify which computation is employed by animals to deal with asynchronous and deceiving information. Learning in ambiguous conditions is in fact an ubiquitous type of neural learning observed in mammals as well as in simpler neural systems (Brembs, 2003) as that of the invertebrate Aplysia (Brembs et al, 2002) or the honey bee (Hammer and Menzel, 1995; Menzel and Müller, 1996; Gil et al, 2007).

When environments are ambiguous due to delayed rewards, or due to concomitant stimuli and actions, the only possibility of finding true cause-effect relationships is to observe repeated occurrences of a reward. By doing that, it is possible to assess the probability of certain stimuli and actions to be the cause of the observed reward. Previous neural models, e.g. (Izhikevich, 2007; Frémaux et al, 2010; Friedrich et al, 2011; Soltoggio and Steil, 2013), solve the distal reward problem applying small weight changes whenever an event indicates an increased or decreased probability of particular pathways to be associated with a reward. With a sufficiently low learning rate, and after repeated reward episodes, the reward-inducing synapses grow large, while all other synapses sometimes increase and sometimes decrease their weights. Those approaches may perform well in reward maximization tasks, but they also cause deterioration of synaptic values because the whole modulated network constantly undergoes synaptic changes across non-reward-inducing synapses. For this reason, only limited information, i.e. those stimulus-action pairs that are frequently rewarded, can be retained even in large networks because the connectivity is constantly rewritten. Interestingly, the degradation of synapses occurs also as a consequence of spontaneous activity as described in Fusi et al (2005). In general, continuous learning, or synapses that are always plastic, pose a treat to previously acquired memory (Senn and Fusi, 2005; Fusi and Senn, 2006; Leibold and Kempter, 2008). Delayed rewards worsen the problem because they amplify synaptic changes caused by reward-unrelated activity. While learning with delayed rewards, current models suffer particularly from the so called *plasticity-stability* dilemma, and *catastrophic forgetting* (Grossberg, 1988; Robins, 1995; Abraham and Robins, 2005).

Synapses may be either coincidentally or causally active before reward deliveries, but which of the two cases applies is unknown due to the ambiguity introduced by delays. How can a system solve this apparent dilemma, and correctly update reward-inducing weights

and leaving the others unchanged? The novel idea in this study is a distinction between two components of a synaptic weight—a volatile component and a consolidated component. Such as distinction is not new in connectionist models (Hinton and Plaut, 1987; Schmidhuber, 1992; Levy and Bairaktaris, 1995; Tieleman and Hinton, 2009; Bullinaria, 2009), however, in the proposed study the idea is extended to model hypothesis testing and memory consolidation with distal rewards. The volatile (or transient) component of the weight may increase or decrease at each reward delivery without immediately affecting the long-term component. It decays over time, and for this reason may be seen as a particular form of short-term plasticity. In the rest of the paper, the terms *volatile*, *transient* and *short-term* are used as synonyms to indicate the component of the weight that decays over time. In contrast, *consolidated*, *long-term*, or *stable* are adjectives used to refer to the component of the weight that does not decay over time.

Short-term volatile weights are *hypotheses* of how likely stimulus-action pairs lead to future rewards. If not confirmed by repeated disambiguating instances, short-term weights decay without affecting the long-term configuration of the network. Short-term synaptic weights and the plasticity that regulates them can be interpreted as implementing Bayesian belief (Howson and Urbach, 1989), and the proposed model interpreted as a special case of a learning Bayesian network (Heckerman et al, 1995; Ben-Gal, 2007). Short-term weights that grow large are therefore those that consistently trigger a reward. The idea in this study is to perform a parsimonious consolidation of weights that have grown large due to repeated and consistent reward-driven potentiation. Such dynamics lead to a consolidation of weights representing established hypotheses.

The novelty of the model consists in implementing dynamics to test *temporal causal hypotheses* with a transient component of the synaptic weight. Transient weights are increased when the evidence suggests an increased probability of being associated with a future reward. As opposed to Izhikevich (2007), in which a baseline modulation results in a weak Hebbian plasticity in absence of reward, in the current model an anti-Hebbian mechanism leads transient weights to be depressed when the evidence suggests no causal relations to future rewards. The distinction between short and long-term components of a weight allows for an implicit estimation of the probability of a weight to be associated with a reward without changing its long-term consolidated value. When coincidental firing leads to an association, which is however not followed by validating future episodes, long-term weight components remain unchanged. The novel plasticity suggests a nonlinear

mechanism of consolidation of a hypothesis in established knowledge during distal reward learning. Thus, the proposed plasticity rule is named Hypothesis Testing Plasticity (HTP).

The current model uses eligibility traces with a decay in the order of seconds to bridge stimuli, actions, and rewards. As it will be clarified later, the decay of transient weights acts instead in the order of hours, thereby representing the forgetting of coincidental event-reward sequences that are not confirmed by consistent occurrences. It is important to note that HTP does not replace previous plasticity models of reward learning, it rather complements them with the additional idea of decomposing the weight in two components, one for hypothesis testing, and one for long-term storage of established associations.

In short, HTP enacts two main principles. The first is monitoring correlations by means of short-term weights and actively pursuing exploration of probably rewarding stimulus-action pairs; the monitoring (or hypothesis evaluation) is done without affecting the long-term state of the network. The second principle is that of selecting few established relationships to be consolidated in long-term stable memory.

HTP is a meta-plasticity scheme and is general to both spiking and rate-based codes. The rule expresses a new theory to cope with multiple rewards, to learn faster and preserve memories of one task in the long term also while learning or performing in other tasks.

2 Method

This section describes the learning problem, overviews existing plasticity models that solve the distal reward problem, and introduces the novel meta-plasticity rule called Hypothesis Testing Plasticity (HTP).

2.1 Operant learning with asynchronous and distal rewards

A newly born learning agent, when it starts to experience a flow of stimuli and to perform actions, has no knowledge of the meaning of inputs, nor of the consequences of actions. The learning process considered here aims at understanding what reward relationships exist between stimuli and actions.

The overlapping of stimuli and actions represents the coexistence of a flow of stimuli with a flow of actions. Stimuli and actions are asynchronous and initially unrelated. The execution of actions is initially driven by internal dynamics, e.g. driven by noise, because the agent's knowledge is a tabula rasa, i.e. is un-
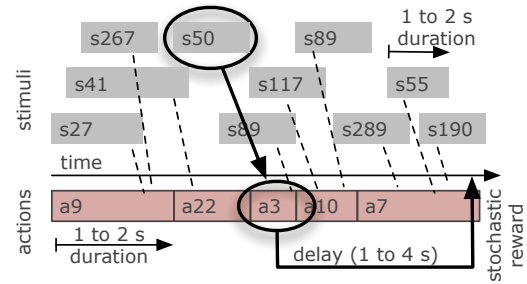
Fig. 1 Graphical representation of the asynchronous flow of stimuli and actions with delayed rewards. The agent perceives an input flow and performs actions. The delayed rewards, as well as the simultaneous presence of stimuli and actions, cause ambiguity as to which stimulus-action pair is the real cause of a reward. The occurrence of more rewards is the only possibility to disambiguate those relationships and discover the invariant causes. In this graphical illustration, when the reward is delivered, 10 stimulus-action pairs (indicated with dash lines) were active in the recent past. All those pairs may be potentially the cause of the reward: in effect, it is only the pair s50-a3 that caused the reward.

biased and agnostic of the world. Spontaneous action generation is a form of exploration. A graphical representation of the input-output flow is given in Fig. 1.

In the setup of the current experiments, at any moment there might be between zero and three stimuli. Stimuli and actions have a random duration between 1 and 2 s. Some actions, if performed when particular stimuli are present, cause the delivery of a global noisy signal later in time (between 1 and 4 s later), which can be seen as a reward, or simply as an unconditioned stimulus. The global reward signal is highly stochastic in the sense that both the delay and the intensity are variable. In the present setting, 300 different stimuli may be perceived at random times. The agent can perform 30 different actions, and the total number of stimulus-action pairs is 9000. The task is to learn which action to perform when particular stimuli are present to obtain a reward.

It is important to note that the ambiguity as to which pairs cause a reward emerges from both the simultaneous presence of more stimuli, and from the delay of a following reward. From a qualitative point of view, whether distinct stimulus-action pairs occurred simultaneously or in sequence has no consequence: a learning mechanism must take into consideration that a set of pairs were active in the recent past. Accordingly, the word *ambiguity* in this study refers to the fact that, at the moment of a reward delivery, several stimulus-action pairs were active in the recent past, and all of them may potentially be the cause of the reward.

2.2 Previous models with synaptic eligibility traces

In simple neural models, the neural activity that triggers an action, either randomly or elicited by a particular stimulus, is gone when a reward is delivered seconds later. For this reason, standard modulated plasticity rules, e.g. (Montague et al, 1995; Soltoggio and Stanley, 2012), fail unless reward is simultaneous with the stimuli. If the reward is not simultaneous with its causes, *eligibility traces* or *synaptic tags* have been proposed as means to bridge the temporal gap (Frey and Morris, 1997; Wang et al, 2000; Sarkisov and Wang, 2008; Päpper et al, 2011).

Previous models with reward-modulated Hebbian plasticity and eligibility traces were shown to associate past events with following rewards, both in spiking models with spike-timing-dependent plasticity (STDP) (Izhikevich, 2007) and in rate-based models with Rarely Correlating Hebbian Plasticity (RCHP) (Soltoggio and Steil, 2013; Soltoggio et al, 2013a). RCHP is a filtered Hebbian rule that detects only highly correlating and highly decorrelating activity by means of two thresholds (see Appendix 2): the effect is that of representing sparse (or rare) spiking coincidence also in rate-based models. RCHP was shown in Soltoggio and Steil (2013) to have computationally equivalent learning to the spiking rule (R-STDP) in Izhikevich (2007).

Spike coincidence in Izhikevich (2007), or highly correlating activity in Soltoggio and Steil (2013), increase synapse-specific eligibility traces. Even with fast network activity (in the millisecond time scale), eligibility traces can last several seconds: when a reward occurs seconds later, it multiplies those traces and reinforces synapses that were active in a recent time window. Given a presynaptic neuron j and a postsynaptic neuron i, the changes of weights w_{ji}, modulation m, and eligibility traces E_{ji}, are governed by

$$\dot{m}(t) = -m(t)/\tau_m + \lambda \cdot r(t) + b \tag{1}$$

$$\dot{w}_{ji}(t) = m(t) \cdot E_{ji}(t) \quad , \tag{2}$$

$$\dot{E}_{ji} = -E_{ji}/\tau_E + \Theta_{ji}(t) \tag{3}$$

where the modulatory signal $m(t)$ is a leaky integrator of the global reward signal $r(t)$ with a bias b; τ_E and τ_m are the time constants of the eligibility traces and modulatory signal; λ is a learning rate. The signal $r(t)$ is the reward determined by the environment. The modulatory signal $m(t)$, loosely representing dopaminergic activity, decays relatively quickly with a time constant $\tau_m = 0.1$ s as measured in Wighmann and Zimmerman (1990); Garris et al (1994). In effect, Eq. (1) is a rapidly decaying leaky integrator of instantaneous reward signals received from the environment. The synaptic trace E is a leaky integrator of correlation episodes Θ. In

Izhikevich (2007), Θ is the STDP(t) function; in Soltoggio and Steil (2013), Θ is implemented by the rate-based Rarely Correlating Hebbian Plasticity (RCHP) that was shown to lead to the same neural learning dynamics of the spiking model in Izhikevich (2007). RCHP is a thresholded Hebbian rule expressed as

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

where α and β are two positive learning rates for correlating and decorrelating synapses respectively, $v(t)$ is the neural output, t_{pt} is the propagation time of the signal from the presynaptic to the postsynaptic neuron, and θ_{hi} and θ_{lo} are the thresholds that detect highly correlating and highly decorrelating activities. RCHP is a nonlinear filter on the basic Hebbian rule that ignores most correlations. Note that the propagation time t_{pt} in the Hebbian term implies that the product is not between simultaneous presynaptic and postsynaptic activity, but between presynaptic activity and postsynaptic activity when the signal has reached the postsynaptic neuron. This type of computation attempts to capture the effect of a presynaptic neuron on the postsynaptic neuron, i.e. the causal pre-before-post situation (Gerstner, 2010), considered to be the link between the Hebb's postulate and STDP (Kempter et al, 1999). The regulation of the adaptive threshold is described in the Appendix 2s. A baseline modulation b can be set to a small value and has the function of maintaining a small level of plasticity.

The idea behind RCHP, which reproduces with rate-based models the dynamics of R-STDP, is that eligibility traces must be created parsimoniously (with rare correlations). When this criterion is respected, both spiking and rate-based models display similar learning dynamics.

In the current model, the neural state u_i and output v_i of a neuron i are computed with a standard rate-based model expressed by

$$u_i(t) = \sum_j (w_{ji} \cdot v_j(t)) + I_i \tag{5}$$

$$v_i(t + \Delta t) = \begin{cases} \tanh(\gamma \cdot u_i(t)) + \xi_i(t) & \text{if } u_i \geq 0 \\ \xi_i(t) & \text{if } u_i < 0 \end{cases} \tag{6}$$

where w_{ji} is the connection weight from a presynaptic neuron j to a postsynaptic neuron i; γ is a gain parameter set to 0.5; $\xi_i(t)$ is a Gaussian noise source with standard deviation 0.02. The input current I is set to 10 when an input is delivered to a neuron. The sampling time is set to 100 ms, which is also assumed to be the propagation time t_{pt} (Eq. (4)) of signals among neurons.

2.3 Hypothesis Testing Plasticity (HTP)

The dynamics of Eqs. (2-3) erode existing synapses because the spontaneous network activity during reward episodes causes synaptic correlations and weight changes. The deterioration is not only caused by endogenous network activity, but it is also caused by the ambiguous information flow (Fig. 1). In fact, many synapses are often increased or decreased because the corresponding stimulus-action pair is coincidentally active shortly before a reward delivery. Therefore, even if the network was internally silent, i.e. there was no spontaneous activity, the continuous flow of inputs and outputs generates correlations that are transformed in weight changes when rewards occur. Such changes are important because they test hypotheses. Unfortunately, if applied directly to the weights, they will eventually wear out existing topologies.

To avoid this problem, the algorithm proposed in this study explicitly assigns the fluctuating dynamics of Eq. (2) to a transient component of the weight. As opposed to the long-term component, the transient component decays over time. Assume, e.g., that one particular synapse had pre and postsynaptic correlating activity just before a reward delivery, but it is not known whether there is a causal relation to the delivery of such a reward, or whether such a correlation was only coincidental. Eq. (2) increases correctly the weight of that synapse because there is no way at this stage to know whether the relation is causal or coincidental. In the variation proposed here, such a weight increase has a short-term nature because it does not represent the acquisition of established knowledge, but it rather represents the increase of probability that such a synapse is related to a reward delivery. Accordingly, weight changes in Eq. (2) are newly interpreted as changes with short-term dynamics

$$\dot{w}_{ji}^{st}(t) = -w_{ji}^{st}/\tau_{st} + m(t) \cdot E_{ji}(t) \qquad (7)$$

where w^{st} is now a transient component of the weight, and τ_{st} is the corresponding decay time constant. The time constant of short-term memory τ_{st} is set to 8 h. In biological studies, short-term plasticity is considered only for potentiation lasting up to 10 minutes (Zucker, 1989; Fisher et al, 1997; Zucker and Regehr, 2002). However, in the idea of this study, the duration of volatile weights represents the duration of an *hypothesis* rather than a specific biological decay. Thus, the value of τ_{st} can be chosen in a large range. A brief time constant ensures that weights decay quickly if rewards are not delivered. This helps maintain low weights but, if rewards are sparse in time, hypotheses are forgotten too quickly. With sporadic rewards, a longer decay may help preserve hypotheses longer in time. If τ_{st} is

set to large values, hypotheses remain valid for an arbitrary long time. This point indicates that, in the current model, short-term weights are intended primarily as *probabilities* of relationships to be true, rather than simply short *time spans* of certain information.

If a stimulus-action pair is active at a particular point in time, but no reward follows within a given interval (1 to 4 s), it would make sense to infer that such a stimulus-action pair is unlikely to cause a reward. This idea is implemented in HTP by setting the baseline modulation value b in Eq. (2) to a small negative value. The effect is that of establishing weak anti-Hebbian dynamics across the network in absence of rewards. Such a setting is in contrast to Izhikevich (2007) in which the baseline modulation is positive. By introducing a small negative baseline modulation, the activation of a stimulus-action pair, and the consequent increase of E, results in a net weight decrement if no reward follows. In other words, high eligibility traces that are not followed by a reward cause a small weight decrease. This modification that decreases a weight if reward does not follow is a core principle in the hypothesis testing mechanism introduced by HTP. By introducing this idea, weights do not need to be randomly depressed by decorrelations, which therefore are not included in the current model.

Finally, the principles of HTP illustrated above can be applied to a reward-modulated plasticity rule such as R-STDP, RCHP, or any rule capable of computing sparse correlations Θ in the neural activity, and consequently E, in Eq. 3. In the current study, a rate-based model plus RCHP are employed. In particular, a simplified version of the RCHP, without decorrelations, is expressed as

$$\Theta_{ji} = \text{RCHP}_{ji}^{+}(t) = +1 \quad \text{if } v_j(t-t_{pt}) \cdot v_i(t) > \theta_{hi} \quad (8)$$

and 0 otherwise (compare with Eq. (4)). Decorrelations may be nevertheless modelled to introduce weight competition[1].

The overall HTP synaptic weight W is the sum of the short-term and long-term components

$$W_{ji}(t) = w_{ji}^{st}(t) + w_{ji}^{lt}(t) \quad . \qquad (9)$$

As the transient component is also contributing to the overall weight, short-term changes also influences how presynaptic neurons affect postsynaptic neurons, thereby biasing exploration policies as it will be explained in the result section.

The proposed model consolidates transient weights in long-term weights when the transient values grow

---

[1] In that case, is essential that the traces E are bound to positive values: negative traces that multiply with the negative baseline modulation would lead to unwanted weight increase.

large. Such a growth indicates a high probability that the activity across that synapse is involved in triggering following rewards. In other words, after sufficient trials have disambiguated the uncertainty introduced by the delayed rewards, a nonlinear mechanism convert hypotheses to certainties. Previous models (Izhikevich, 2007; O'Brien and Srinivasan, 2013; Soltoggio and Steil, 2013) show a *separation* the of weight values between reward-inducing synapses (high values) and other synapses (low values). In the current model, such a separation is exploited and identified by a threshold Ψ loosely set to a high value, in this particular setting to 0.95 (with weights ranging in [0, 1]). The conversion is formally expressed as

$$\dot{w}_{ji}^{lt}(t) = \rho \cdot H(w_{ji}^{st}(t) - \Psi) \quad , \tag{10}$$

where H is the Heaviside function and ρ is a consolidation rate, here set to 1/1800 s. Note that in this formulation, $\dot{w}_{ji}^{lt}(t)$ can only be positive, i.e. long-term weights can only increase: a variation of the model is discussed later and proposed in the Appendix 1. The consolidation rate ρ means that short-term components are consolidated in long-term components in half an hour when they are larger than the threshold Ψ. A one-step instantaneous consolidation (less biologically plausible) was also tested and gave similar results, indicating that the consolidation rate is not crucial.

The threshold Ψ represents the point at which an hypothesis is considered true, and therefore consolidated in long-term weight. The idea is that, if a particular stimulus-action pair has been active many times consistently before a reward, such stimulus-action pair is indeed causing the reward. Interestingly, because the learning problem is inductive and processes are stochastic, certainty can never be reached from a purely theoretical view point. Assume for example that, on average, every second one reward episode occurs with probability $p = 10^{-2}$ and leads short-term weights that were active shortly before the delivery to grow of 0.05[2]. To grow to saturation, a null weight needs 1) to be active approximately 20 times before reward deliveries and 2) not to be active when rewards are not delivered. If a synapse is not involved in reward delivery, the probability of such a synapse to reach Ψ might be very low in the oder of p^{20}, i.e. 10^{-40}. The complex and non-stationary nature of the problem does not allow for a precise mathematical derivation. Such a probability is in fact affected by a variety of environmental and network factors such as the frequency and amount of reward, the total number of stimulus-action pairs, the firing

rate of a given connection, the number of intervening events between cause and effect (reward), and the contribution of the weight itself to a more frequent firing. Nevertheless, previous mathematical and neural models that solve the distal reward problem rely on the fact that consistent relationships occurs indeed consistently and more frequently than random events. As a consequence, after a number of reward episodes, the weight that is the true cause of reward has been *accredited* (increased) more than any other weight. The emergence of a *separation* between reward-inducing weights and other weights is observed in Izhikevich (2007); O'Brien and Srinivasan (2013); Soltoggio and Steil (2013). The proposed rule exploits this separation between reward-inducing and non-reward-inducing synapses to consolidate established relationship in long-term memory. The dynamics of Eqs. (7-10) are referred to as Hypothesis Testing Plasticity (HTP).

The long-term component, once is consolidated, cannot be undone in the present model. However, reversal learning can be easily implemented by adding complementary dynamics that undo long-term weights if short-term weights become heavily depressed. Such an extension is proposed in the Appendix 1.

The role of short-term plasticity in improving reward-modulated STDP is also analyzed in a recent study by O'Brien and Srinivasan (2013). With respect to O'Brien and Srinivasan (2013), the idea in the current model is general both to spiking and rate-based coding and is intended to suggest a role of short-term plasticity rather than to model precise biological dynamics. Moreover, it does not employ reward predictors, it focuses on the functional roles of long-term and short-term plasticity, and does not necessitate the Attenuated Reward Gating (ARG).

Building on models such as Izhikevich (2007); Florian (2007); Friedrich et al (2011); Soltoggio and Steil (2013), the current model introduces the concept of testing hypotheses with ambiguous information flow. The novel meta-plasticity model illustrates how the careful promotion of weights to a long-term state allows for retention of memory also while learning new tasks.

2.4 Action selection

Action selection is performed by initiating the action corresponding to the output neuron with the highest activity. Initially, selection is mainly driven by neural noise, but as weights increase, the synaptic strengths bias action selection towards output neurons with strong incoming connections. One action has a random duration between 1 and 2 s. During this time, the action feeds back to the output neuron a signal $I = 0.5$. Such

---

[2] The exact increment depends on the learning rate, on the exact circumstantial delay between activity and reward, and on the intensity of the stochastic reward.
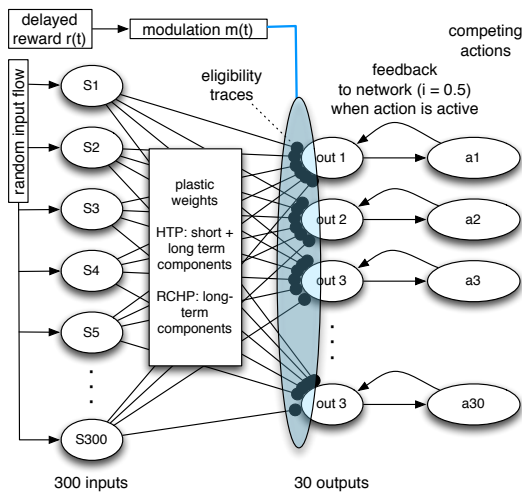
Fig. 2 Graphical representation of the feed-forward neural network for distal reward learning tested with both the basic RCHP and the novel HTP. One stimulus is delivered to the network by activating a corresponding input neuron. Each weight is plastic and has a trace associated. The modulatory signal is an additional input that modulates the plasticity of all the weights. The sampling time is 100 ms, but the longer temporal dynamics given by delayed rewards is captured by the 4 s time constant of the eligibility traces. The output neuron with the highest activity initiates an action. The action then feeds back to that neuron a feedback signal which helps input and output correlate correctly (see Appendix).

a signal is important to make the *winning* output neuron "aware" that it has triggered an action. Computationally, the feedback to the output neuron increases its activity, thereby inducing correlations on that particular input-output pair, and causing the creation of a trace on that particular synapse. Feedback signals to output neurons are demonstrated to help learning also in Urbanczik and Senn (2009); Soltoggio et al (2013a). The overall structure of the network is graphically represented in Fig. 2.

Further implementation details are in the Appendix. The Matlab code used to produce the results is made available as support material.

3 Results

In this section, simulation results present the computational properties of HTP. A first test is a computational assessment of the extent of weight unwanted change due to distal rewards when one single weight component is used. The learning and memory dynamics of the novel plasticity are tested with the network of Fig. 2 on a set of learning scenarios. The dynamics of HTP are illustrated in comparison to those of the single weight component implemented by the basic RCHP.

3.1 Weight deterioration and stochasticity with distal rewards

Algorithms that solve the distal reward problem have so far focused on reward maximization (Urbanczik and Senn, 2009; Frémaux et al, 2010; Friedrich et al, 2011). Little attention was given to non-reward-inducing weights. However, non-reward-inducing weights are often the large majority of weights in a network. Their changes are relevant to understand how the whole network evolves over time, and how memory is preserved (Senn and Fusi, 2005). The test in this section analyzes the side effects of distal rewards on non-reward-inducing weights.

Assume that a correlating event between two neurons across one synapse σ represents a stimulus-action pair that is *not* causing a reward. Due to distal rewards, the synapse σ might occasionally register correlation episodes in the time between the real cause and a delayed reward: that is in the nature of the distal reward problem. All synapses that were active shortly before a reward might be potentially the cause, and the assumption is that the network does not know which synapse (or set of synapses) are responsible for the reward (thus the whole network is modulated).

The simulation of this section is a basic evaluation of a weight updating process. The term $m(t) \cdot E(t)$, which affects Eqs. (2) and (7), and expresses a credit assignment, is pre-determined according to different stochastic regimes. The purpose is to evaluate the difference between single-weight-component and two weight-component dynamics illustrated by Eqs. (9) and (10), independently of specific reward-learning plasticity rule.

The value of a weight σ is monitored each time an update occurs. Let us assume arbitrarily that a correlation across σ and a following unrelated reward occurs coincidentally every five minutes. Three cases are considered. In phase one, the weight is active coincidentally before reward episodes (i.e. no correlation with the reward). For this reason, modulation causes sometimes increments and sometimes decrements. Such setting represents algorithms that do not have an "unsupervised bias", e.g. Urbanczik and Senn (2009); Frémaux et al (2010), which guarantee that the reward maximization function has a null gradient if the weight does not cause a reward. To reproduce this condition here, the stochastic updates in phase 1 have an expected value of zero. In a second phase, weight updates cease to occur, representing the fact that the weight σ is never active before rewards (no ambiguity in the information flow). In a third phase, the weight σ is active before rewards more often than not, i.e. it is now mildly correlated to reward episodes, but in a highly stochastic regime.
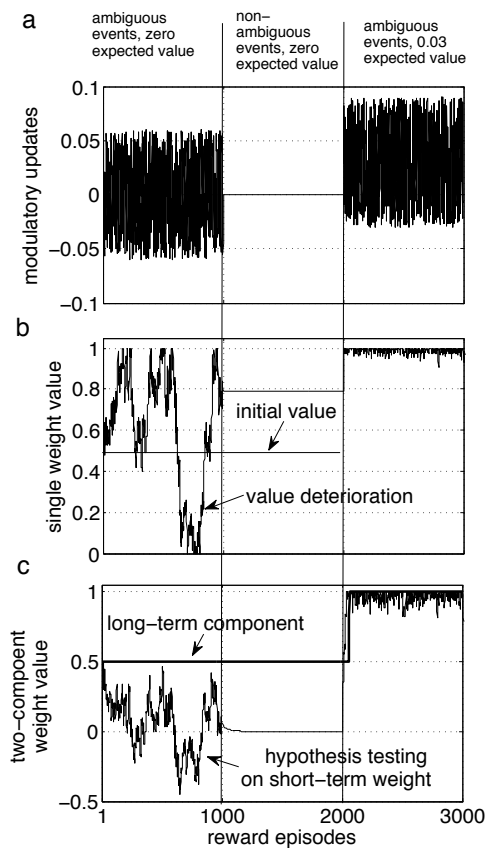
Fig. 3 Illustration of one versus two weight component dynamics with ambiguous updates due to distal rewards. (a) Random values of weight updates illustrate three cases: from 1 to 1000 the weight is not related to a reward, the updates have an expected value of zero. From 1001 to 2000, the weight is never active when rewards occur, there are no updates. From 2001 to 3000, random updates have non-zero mean, i.e. the activity of the weight is correlated to a reward signal. (b) Changes in the weight σ when σ is composed of one single component. The initial value of 0.5 is progressively forgotten. (c) Changes in the weight of σ when σ is composed of two components. The initial value of 0.5 is preserved by the long-term component, while the short-term component acts as a monitor for correlations with the reward signal (correctly detected after step 2000).

Fig. 3a illustrates weight updates that were randomly generated and draw from the distributions U(-0.06,0.06) for the reward episodes 1 to 1000, U(0,0) for the reward episodes from 1001 to 2000, and U(-0.03,0.09) for the reward episodes 2001 to 3000. The distribution in the first 1000 reward episodes represents a random signal with an expected value of zero, i.e. the weight σ is not associated with the reward. Figs. 3bc show respectively the behaviors of a single-weight-component rule and of a two-weight-component rule with weight decay on the short-term component. In the single-weight-component case (Fig. 3b), despite the updates have an expected value of zero, the weight loses its original value of 0.5.

The initial value of 0.5 is chosen arbitrarily to be in between 0 and 1 to observe both positive and negative variations from its original value. The forgetting of the original value of σ is logical because even if the expected value of the updates is zero, there is no mechanism to "remember" its initial value. The weight undergoes a random walk, or diffusion, that leads to information loss. The example in Fig. 3b shows that the weight change is not negligible, ranging from 0 to saturation. Note that the rate of change, and the difference between the original value and the final value in this example is only illustrative. In a neural network, updates are a function of more variables including the strength of the synapse itself and the neural activity. However, the current example captures an important aspects of learning with delayed rewards: regardless of the plasticity rule, coincidental events in a neural network may lead to unwanted changes. The example is useful to show that a plasticity rule with a single weight-component, even if not affected by the "unsupervised bias", disrupts existing weights that are not related to rewards but are active before rewards. Fig. 3c instead shows that a two-weight-component rule preserves its long-term component, while the short-term component is affected by the random updates. However, due to its decay, the short-term component tends to return to low values if the updates have limited amplitude and an expected value of zero. If rewards and activity across σ never occur together (reward episodes from 1001 to 2000), there is no ambiguity and σ is clearly not related to rewards: the single-weight-component rule maintains the value of σ, while the two-weight-component rule has a decay to zero of the short-term component. Finally, in the phase from reward episode 2001 to 3000, the updates have a positive average sign, but are highly stochastic: both rules bring the weight to its saturation value 1. In particular, the two-weight-component rule brings the long-term component to saturation as a consequence of the short-term component being above the threshold level.

This simple computational example, which does not yet involve a neural model, shows that distal reward learning with a single-weight component leads to weight deterioration of currently non-reward-inducing weights. A two-weight-component rule instead has the potential of preserving the values of weights in the long-term component, while simultaneously monitoring the correlation to reward signals by means of the short-term component. The principle illustrated in this section is used by HTP on a neural model with the results presented in the following sections.

3.2 Learning without forgetting

Three different learning scenarios are devised to test the neural learning with the network in Fig. 2. Each learning scenario lasts 24 h of simulated time and rewards 10 particular stimulus-action pairs (out of a total of 9000 pairs). A scenario may be seen as a learning task composed of 10 subtasks (i.e. 10 stimulus-action pairs). The aim is to show the capability of the plasticity rule to learn and memorize stimulus-action pairs across multiple scenarios. Note that the plasticity rule is expected to bring to a maximum value all synapses that represent reward-inducing pairs (Fig. 2).

The network was simulated in scenario 1 (for 24 h), then in scenario 2 (additional 24 h), and finally in scenario 3 (again 24h). During the first 24 h (scenario 1), the rewarding input-output pairs are chosen arbitrarily to be those with indices (i, i) with $1 \leq i \leq 10$. When a rewarding pair occurs, the input $r(t)$ (normally 0) is set to 0.5 ± 0.25 at time $t + \varphi$ with φ drawn from a uniform distribution $U(1, 4)$. φ represents the delay of the reward. With this setting, not only is a reward occurring with a random variable delay, but its intensity is also random, making the solution of the problem even more challenging. In the second scenario, the rewarding input-output pairs are $(i, i-5)$ with $11 \leq i \leq 20$. No reward is delivered when other stimulus-action pairs are active. A third scenario has again different rewarding pairs as summarized in Table 1. The arbitrary stimulus-action rewarding pairs were chosen to be easily seen on the weight matrix as diagonal patterns. While stimuli in the interval 31 to 300 occur in all scenarios, stimuli 1 to 10 occur only scenario 1, stimuli 11 to 20 in scenario 2 and stimuli 21 to 30 in scenario 3. This setting is meant to represents the fact that the stimuli that characterize rewards in one scenario are not present in other scenarios, otherwise all scenarios would be effectively just one. While in theory it would be possible to learn all relationships simultaneously, such a division in tasks (or scenarios) is intended to test learning, memory and forgetting when performing different tasks at different times. It is also possible to interpret a task as a focused learning session in which only a subset of all relationships are observed.

Fig. 4a shows the cumulative weights of the reward-causing synapses throughout the 72 h of simulation, i.e. scenario 1, followed by scenario 2, followed by scenario 3. RCHP, while learning in the second scenario, causes a progressive forgetting of the knowledge acquired during the first scenario. HTP, when learning in scenario 2, also experiences a partial decay of the weights learned during scenario 1. The partial decay corresponds to the short-term weight components. While

| Scenario	Rewarding stimulus-action pairs	Perceived stimuli
1	(1,1);(2,2)...(10,10)	1 to 10 and 31 to 300
2	(11,6);(12,7)...(20,15)	11 to 20 and 31 to 300
3	(21,1);(22,2)...(30,10)	21 to 300

Table 1 Summary of learning scenarios, rewarding stimulus-action pairs, and pool of perceived stimuli. When one of the listed stimulus-action pair occurs, a stochastic reward drawn from the uniform distribution $U(0.25, 0.75)$ is delivered with a variable delay between 1 and 4 s.

learning in scenario 2, which represents effectively a different environment, the stimuli of scenario 1 are absent, and the short-term components of the relative weights decay to zero. In other words, while learning in scenario 2, the hypotheses on stimulus-action pairs in scenario 1 are forgotten, as in fact hypotheses cannot be tested in the absence of stimuli. However, the long-term components, which were consolidated during learning in scenario 1, are not forgotten while learning in scenario 2. Similarly it happens in scenario 3. These dynamics lead to a final state of the networks shown in Fig. 4b. The weight matrices show that, at the end of the 72 h simulation, RCHP encodes in the weights the reward-inducing synapses of scenario 3, but has forgotten the reward-inducing synapses of scenario 1 and 2. Even with a slower learning rate, RCHP would deteriorate weights that are not currently causing a reward because coincidental correlations and decorrelations alter all weights in the network. In contrast, the long-term component in HTP is immune to single correlation or decorrelation episodes, and thus it is preserved.

Learning without forgetting with distal rewards is for the first time modeled in the current study by introducing the assumption that established relationships in the environments, i.e. long-term weights, are stable and no longer subject to hypothesis evaluation.

3.3 The benefit of memory and the preservation of weights

The distinction between short and long-term weight components was shown in the previous simulation to maintain the memory of scenario 1 while learning in scenario 2, and both scenarios 1 and 2 while learning in scenario 3. One question is whether the preservation of long-term weights is effectively useful when revisiting a previously learned scenario. A second fundamental question in this study is whether all weights, reward-inducing and non-reward-inducing, are effectively preserved. To investigate these two points, the simulation was continued for additional 24 h in which the previously seen scenario 1 was revisited.
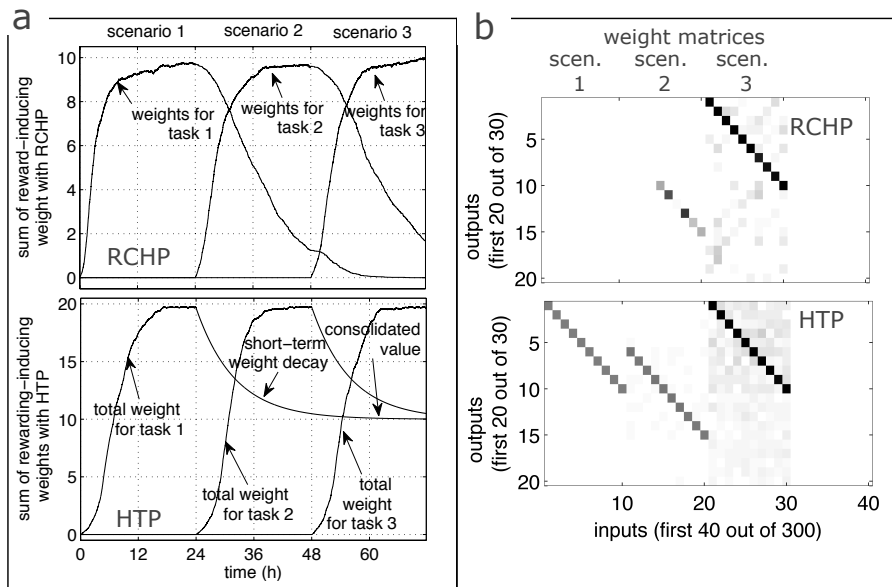
Fig. 4 Learning in three consecutive scenarios. (a) The cumulative total weight of the 10 rewarding synapses (averaged over 10 independent simulations) is shown during the 72 h learning with both RCHP (top graph) and HTP (bottom graph). In the first scenario (first 24 h), the learning leads to a correct potentiation of most reward-inducing synapses. The learning in a second and third scenario with RCHP causes a progressive dismantling of the weights that were reinforced before. HTP identifies consistently all reward-inducing synapses, and does not forget the knowledge of scenario 1 while learning scenario 2 and 3. The partial decay of weights with HTP is due to the short-term component. (b) Partial view of the weight matrix at the end of the 72 h simulation. The view is partial because the overall matrix is 30 by 300: the image only shows the part of the matrix where weights representing relationships are learned. The color represents the strength of the weights, from white (minimum value) to black (maximum value). The high synaptic weights observable under RCHP are those related to scenario 3, because scenarios 1 and 2 are forgotten. The weight matrix with HTP has clearly identified and maintained the 10 rewarding pairs in each scenario.

The utility of memory is shown with the rate of reward per hour as shown in Fig. 5. RCHP performs poorly when scenario 1 is revisited: it re-learns it as if it had never seen it before. HTP instead performs immediately well because the network remembers the stimulus-response pairs in scenario 1 that were learned 72 hours before. Under the present conditions, long-term weights are preserved indefinitely, so that further learning scenarios can be presented to the network without compromising the knowledge acquired previously.

Eq. (10) allows long-term weights to increase, but not to decrease. Therefore, the analysis of weight changes is simplified in the sense that null long-term components at the end of the run are guaranteed not to have experienced any change. Fig. 6 shows the histogram of the long-term synaptic weights after 96 h of simulation with HTP. After hundreds of thousand of stimulus-action pairs, and thousands of reward episodes, none of the 8970 synapses representing non-reward-inducing stimulus-action pairs was modified. Those weights were initially set to zero, and remained so, demonstrating that the stable configuration of the network was not altered during distal reward learning. This fact is remarkable considering that the probability of activation of all

9000 pairs is initially equal, and that many disturbing stimuli and non-rewarding pairs are active each time a delayed reward is delivered. This accuracy and robustness is a direct consequence of the hypothesis testing dynamics in the current model: short-term weights can reach high values, and therefore can be consolidated in long-term weights, only if correlations across those weights are consistently followed by a reward. If not, the long-term component of weights is immune to deterioration and preserves its original value.

3.4 Improved disambiguating capabilities and consequences for learning speed and reliability

An interesting aspect of HTP is that the change of short-term weights also affects the overall weight W in Eq. (9). Thus, an update of w_{st} also changes (although only in the short term) how input signals affect output neurons, thereby also changing the decision policy of the network. Initially, when all weights are low, actions are mainly determined by noise in the neural system (introduced in Eq. (6)). The noise provides an unbiased mechanism to explore the stimulus-action space. As more rewards are delivered, and hypotheses
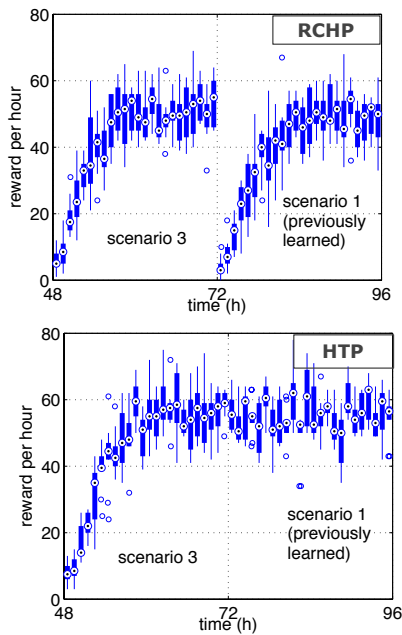
Fig. 5 Amount of reward per hour (box plot statistics over 10 independent trials). RCHP, when revisiting scenario 1, needs to relearn the reward-inducing synapses: those weights were reinforced initially (simulation time 0-24 h), but later at time 72 h, those weights, which where not rewarded, deteriorated and dropped to low values. Although relearning demonstrates the capability of solving the distal reward problem, the network with HTP instead demonstrates that knowledge is preserved and reward rates are immediately high when revisiting scenario 1.
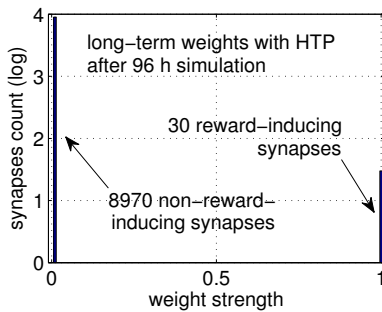

Fig. 6 Histogram of the long-term weights with HTP after the 96 h of simulation, i.e. after performing in scenarios 1, 2, 3 and then 1 again. The long-term components of the weights represent the reward-inducing synapses (an arbitrary set of 30 synapses). All the 8970 non-reward-inducing synapses remain with null weight. This means that the network has not changed any of the weights that are not reward-related. On the other hand, all 30 reward-inducing synapses are identified and correctly consolidated in long-term memory.

are formed (i.e. weights increase), exploration is biased towards stimulus-action pairs that were active in the past before reward delivery. Those pairs include also non-reward-inducing pairs that were active coincidentally, but they certainly include the reward-triggering ones. Such dynamics have two consequences according to whether a reward occurs or not. In the case a reward occurs again, the network will strengthen even more particular weights which are indeed even more likely to be associated with rewards. To the observer, who does not know at which point short-term weights are consolidated in long-term, i.e. when hypotheses are consolidated in certainties, the network acts as if it knows already, although in reality is guessing (and guessing correctly). By doing so, the network actively explores certain stimulus-action pairs that appear "promising" given the past evidence.

The active exploration of a subset of stimulus-action pairs is particularly effective also when a reward fails to occur, i.e. when one hypothesis is false. The negative baseline modulation (term b in Eq. (2)) implies that stimulus-action pairs with high eligibility traces (i.e. that were active in the recent past) but are not followed by rewards decrease their short-term weight components. In a way, the network acts as if trying out potentially reward-causing pairs (pairs whose weight was increased previously), and when rewards do not occur, drops their values, effectively updating the belief by lowering the short-term components of those weights.

What are the consequences of these dynamics? An answer is provided by the weight distribution at the end of learning. The histograms in Fig. 7 show that, in contrast to the single-weight rule (upper histograms), HTP separates clearly the reward-inducing synapses from the others (lower histograms). Such a clear separation is then exploited by HTP by means of the threshold Ψ to consolidate reward-inducing weights. The clear separation also provides an insight onto why HTP appeared so reliable in the present experiments. In contrast, RCHP alone cannot separate synapses very distinctly. Such a lack of separation between reward-inducing and non-reward-inducing weights can also be observed in Izhikevich (2007); O'Brien and Srinivasan (2013). Large synapses in the run with RCHP represent, like for HTP, hypotheses on input-output-reward temporal patterns. However, weights representing false hypotheses are not easily depressed under RCHP or R-STDP that rely only on decorrelations to depress weights. In fact, a large weight causes that synapse to correlate even more frequently, biasing the exploration policy, and making the probability of such an event to occur coincidentally before a reward even higher. Such a limitation in the models in Izhikevich (2007); Florian
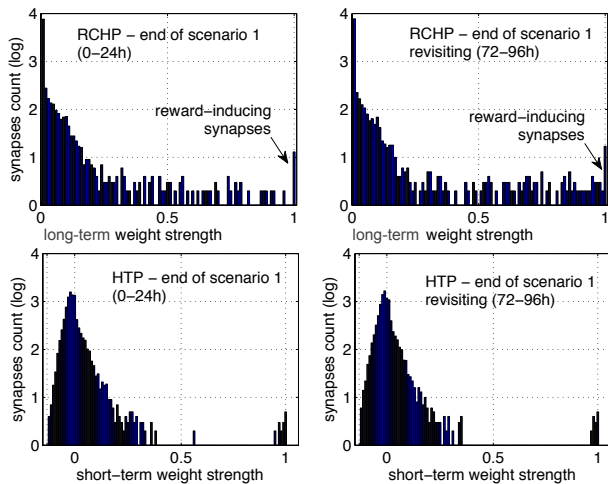
Fig. 7 Histograms of the weight distribution after learning (long-term total weight for RCHP and short-term for HTP). RCHP (upper graphs) does not appear to separate well the reward-inducing synapses from the others. In particular, in the last phase of the simulation (h 72-96, upper right graph), many synapses reach high values. HTP instead (lower graphs) separates distinctly the short-term components of reward-inducing synapses from the others. At the end of the last simulation (96 h, lower right graph), the separation remains as large as it was at 24 h, indicating that such a weight distribution is stable.

(2007); O'Brien and Srinivasan (2013); Soltoggio and Steil (2013) is removed in the current model that instead explicitly depresses synapses that are active but fail to trigger rewards. Note that HTP pushes also some short-term weights below zero. Those are synapses that were active often but no reward followed. In turn, these lower weights are unlikely to trigger actions.

Fig. 7 shows the weight distribution and the separation between reward-inducing and non-reward-inducing synapses at the end of a 4-day simulated time. One might ask whether this separation and distribution is stable throughout the simulation and over a longer simulation time. One additional experiment was performed by running the learning process in scenario 1 for 12 days of simulated time, i.e. an extended amount of time beyond the initial few hours of learning. Fig. 8a shows the average value of the reward-inducing synapses, the average value of non-reward-inducing synapses and the strongest synapse among the non-reward-inducing ones. The consistent separation in weight between synapses that do or do not induce a delayed reward indicates that the value of Ψ, set to 0.95 in all experiments of this study, is not a critical parameter. If the plasticity rule is capable of separating clearly the reward-inducing synapses from the non-reward-inducing synapses, the parameter Ψ can be set to any high value that is unlikely to be reached by non-reward-inducing synapses. Fig. 8b plots the histogram of weight distribution at
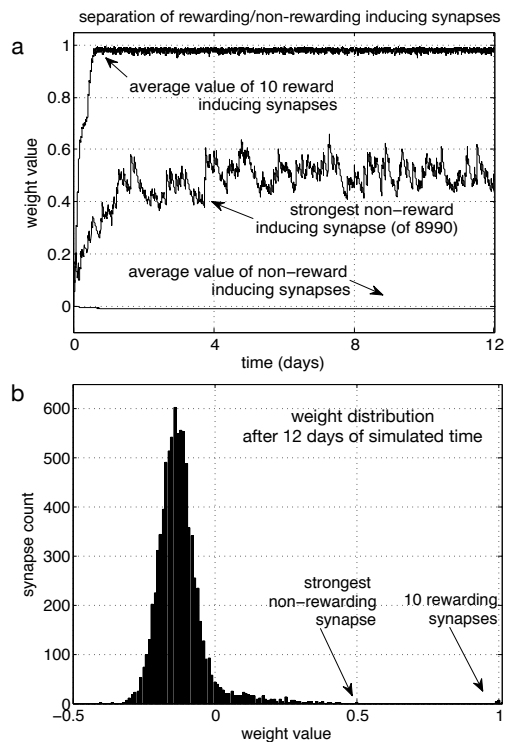
Fig. 8 Separation of reward-inducing and non-reward-inducing synapses monitored during a long simulation for assessing stability. (a) Scenario 1 is simulated for 12 days (288 simulated hours). The plot shows the average value of the 10 reward-inducing synapses, the strongest individual synapse among the other 8 990 non-reward-inducing synapses, and the average value of all non-reward-inducing synapses. (b) At the end of the 12 days of simulation, the distribution of all weights is shown in the histogram. The number of non-reward-inducing synapses that is potentiated beyond the value 0.1 is only 2.1% of the total.

the end of the simulation (after 12 days of simulated time). The histogram shows clearly that although the strongest non-reward-inducing synapses throughout the run oscillates approximately around 0.5, the percentage of non-reward-inducing synapses that are potentiated is very small (only 2% of synapses exceed 0.1 in strength).

The fact that HTP separates more clearly rewarding from non-rewarding weights has a fundamental consequence on the potential speed of learning. In fact, high learning rates in ambiguous environments are often the cause of erroneous learning. If a stimulus-action pair appears coincidentally a few times before a reward, a fast learning rate will increase the weight of this pair to high values, leading to what can be compared to superstitious learning (Skinner, 1948; Ono, 1987). However, if HTP, for the reasons explained above, is capable of better separation between reward-inducing and non-reward-inducing weights, and in particular is capable of depressing false hypotheses, the consequence is that

HTP can adopt a faster learning rate with a decreased risk of superstitious learning.

This section showed that the hypothesis testing rule can improve the quality of learning by (a) biasing the exploration towards stimulus-action pairs that were active before rewards and (b) avoiding the repetition of stimulus-action pairs that in the past did not lead to a reward. In turn, such dynamics cause a clearer separation between reward-inducing synapses and the others, implementing an efficient and potentially faster mechanism to extract cause-effect relationships in a deceiving environment.

3.5 Discovering arbitrary reward patterns

When multiple stimulus-action pairs cause a reward, three cases may occur: 1) each stimulus and each action may be associated to one and only one reward-inducing pair; 2) one action may be activated by more stimuli to obtain a reward; 3) one stimulus may activate different actions to obtain a reward. The cases 1) and 2) were presented in the previous experiments. The case 3) is particular: if more than one action can be activated to obtain a reward, given a certain stimulus, the network may discover one of those actions, and then exploit such pair without learning which other actions also lead to rewards. These dynamics represent an agent who exploits one rewarding action but performs poor exploration, and therefore fails to discover all possible rewarding actions. However, if exploration is enforced occasionally even during exploitation, in the long term the network may discover all actions that lead to a reward given one particular stimulus. To test the capability of the network in this particular case, two new scenarios are devised to reward all pairs identified by a checker board pattern on the weight matrix in a 6 by 12 rectangle, in which each scenario rewards the network that discovers the connectivity pattern of a single 6 by 6 checker board. Each stimulus in the range 1 to 6 in a first scenario, and 7 to 12 in a second scenario, can trigger three different actions to obtain a reward. The two tasks were performed sequentially and lasted each 48 h of simulated time.

A first preliminary test (data not shown), both with RCHP and HTP, revealed that, unsurprisingly, the network discovers one rewarding action for each stimulus and consistently exploits that action to achieve a reward, thereby failing to discover other rewarding actions. Interestingly, such a behavior might be optimal for a reward maximization policy. Nevertheless, a variation of the experiment was attempted to encourage exploration by reducing the neural gain γ in Eq. (6) from 0.5 to 0.1. The neural gain expresses the effect of inputs on output neurons: by reducing it, internal noise might occasionally lead to exploration even when a stimulus is known to lead to a reward with a given action. Because exploration is performed occasionally while the network exploits the already discovered reward-inducing pairs, hypotheses are also tested sporadically, and therefore need to remain alive for a longer time. The time constant τ_{st} of the short-term weight was set in this particular simulation to 24 h. For the same reason, the number of actions was limited to 10, i.e. only 10 output neurons, so that exploration is performed on a slightly reduced search space.

Fig. 9 shows the matrixes of the long-term weights after 96 h of simulated time with RCHP (panel a) and with HTP (panel b). RCHP, as already seen in previous experiments, forgets scenario 1 to learn scenario 2. From the matrix in Fig. 9a it is also evident that RCHP did not increase correctly all weights. Some weights that are non-reward-inducing are nevertheless high. It is remarkable instead that HTP (Fig. 9b) discovers the correct connectivity pattern that not only maximizes the reward, but it also represents all rewarding stimulus-action pairs over the two scenarios. The test shows that HTP remains robust even in conditions in which exploration and exploitation are performed simultaneously. The test demonstrates that if the time-constant of transient weights is sufficiently slow, HTP leads to the discovery of reward-inducing weights even if their exploration is performed sporadically.

4 Discussion

The neural model in this study processes input-output streams characterized by ambiguous stimulus-action-reward relationships. Over many repetitions, it distinguishes between coincidentally and causally related events. The flow is ambiguous because the observation of one single reward does not allow for the unique identification of the stimulus-action pair that caused it. The level of ambiguity can vary according to the environment and can make the problem more difficult to solve. Ambiguity increases typically with the delay of the reward, with the frequency of the reward, with the simultaneous occurrence of stimuli and actions, and with the paucity of stimulus-action pairs. The parameters in the neural model are set to cope with the level of ambiguity of the given input-output flow. For more ambiguous environments, the learning rate can be reduced, resulting in a slower but more reliable learning.

HTP proposes a model in which short-term plasticity does not implement only a duration of a memory (Sandberg et al, 2003), but it rather represents the
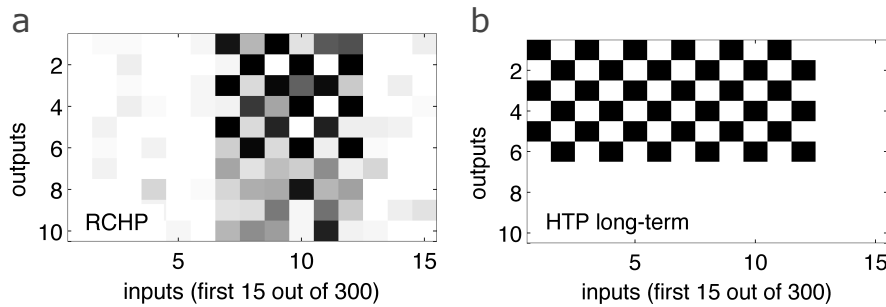
Fig. 9 Learning arbitrary connectivity patterns: partial view of weight matrices. The color represents the strength of the weights, from white (minimum value) to black (maximum value). (a) RCHP attempts to learn a checker board pattern on 12 inputs and 6 outputs in two consecutive scenarios. After 96 h of simulated time, the rule has discovered an approximation of the pattern for the second task (inputs 7 to 12) but has forgotten the first task. The strengths of the weights do not represent very accurately the reward-inducing pairs (compare with panel b). (b) HTP discovers the exact pattern of connectivity that represents reward conditions in the environment across two scenarios that are learned in sequence.

uncertain nature of hypotheses with respect to established facts. Computationally, the advantages of HTP with respect to previous models derive from two features. A first feature is that HTP introduces long-term and short-term components of the weight with different functions: the short-term component tests hypotheses by monitoring correlations; the long-term component consolidates established hypotheses in long-term memory. A second feature is that HTP implements a better exploration: transient weights mean that stimulus-action pairs are *hypotheses* to be tested by means of a targeted exploration of the stimulus-response space.

Previous models, e.g. Izhikevich (2007); Friedrich et al (2011); Soltoggio and Steil (2013), that solved the distal reward problem with one single weight component, cannot store information in the long term unless those weights are frequently rewarded. In contrast, HTP consolidates established associations to long-term weights. In this respect, any R-STDP-like learning rule can learn current reward-inducing relationships, but will forget those associations if the network is occupied in learning other tasks. HTP can build up knowledge incrementally by preserving neural weights that have been established to represent correct associations. HTP is the first rule to model incremental acquisition of knowledge with highly uncertain cause-effect relationships due to delayed rewards.

As opposed to most reward modulated plasticity models, e.g. (Legenstein et al, 2010; O'Brien and Srinivasan, 2013), the current network is modulated with raw reward signals. There is not an external value storing expected rewards for a given stimulus-action pair. Such reward predictors are often additional computational or memory units outside the network that help plasticity to work. The current model instead performs all computation within the network. In effect, expected

rewards are computed implicitly, and at the end very accurately, by the synaptic weights themselves. In fact, the synaptic weights, representing an indication of the probability of a future reward, do also implicitly represent the expected reward of a given stimulus-action pair. For example, a synaptic weight that was consolidated in long-term weight represents the high expectation of a future reward. The weight matrix in Fig. 4b (bottom matrix) is an accurate predictor of all rewarding pairs (30) across three different scenarios.

The last experiment showed that the novel plasticity rule can perform well under highly explorative regimes. As opposed to rules with a single weight component, HTP is capable of both maintaining strong weights for exploiting reward conditions, and exploring new stimulus-action pairs. By imposing an arbitrary set of reward-inducing pairs, e.g. the environmental reward conditions are expressed by a checker board on the weight matrix, the last experiment showed that HTP can use very effectively the memory capacity of the network.

The model can also be seen as a high-level abstraction of memory consolidation (McGaugh, 2000; Bailey et al, 2000; Lamprecht and LeDoux, 2004; Dudai, 2004; Mayford et al, 2012) under the effect of delayed dopaminergic activity (Jay, 2003), particularly at the synaptic level as the transition from early-phase to late-phase LTP (Lynch, 2004; Clopath et al, 2008). The consolidation process, in particular, expresses a meta-plasticity mechanism (Abraham and Bear, 1996; Abraham and Robins, 2005; Abraham, 2008), with similarities to the cascade model in Fusi et al (2005), because frequent short-term updates are preconditions for further long-term potentiation (Goelet et al, 1986; Nguyen et al, 1994). By exploiting synaptic plasticity with two different timescales (short and long-term), the current model also contributes to validating the growing view that

multiple timescale plasticity is beneficial in a number of learning and memory models (Abbott and Regehr, 2004; Fusi et al, 2005, 2007). The dynamics presented in this study do not reproduce or model biological phenomena (Zucker and Regehr, 2002). Nevertheless, this computational model proposes a link between short-term plasticity and short-term memory, suggesting the utility of fading short-term memories (Jonides et al, 2008), which may not be a shortcoming of neural systems, but rather a useful computational tool to distinguish between coincidental and reoccurring events.

It is interesting to ask which conditions may lead HTP to fail. HTP focuses on and exploits dynamics of previously proposed reward learning rules that aim at separating rewarding pathways from other non-rewarding pathways. Such a separation is not always easy to achieve. For example, in a plot in Izhikevich (2007) (Fig. 1d), a histogram of all weights shows that the separation between the rewarding synapse and all other synapses is visible but not large. The original RCHP, as reproduced in this study, may also encounter difficulties in creating clear separations as shown in Fig. 7. In short, HTP prescribes mechanisms to create a clear separation between reward-inducing and non-reward-inducing synapses: if this cannot be achieved, HTP cannot be used to consolidate long-term weights. This may be the case when the network is flooded with high levels of reward signals. As a general rule, whenever the input-output flow is ambiguous, plasticity rules require time to separate rewarding weights from non-rewarding weights. A fast learning rate is often the cause of failure. Interestingly, a fast learning rate with distal rewards can be imagined as a form of superstitious type of learning, in which conclusions are drawn from few occurrences of rewards (Skinner, 1948; Ono, 1987).

If learning rates are small (or similarly if rewards are small in magnitude), would not the decay of transient weights in HTP prevent learning? The answer is that the decay of the transient weights, in this study set to 8h (or 24h for the last experiment), represents the time of one learning scenario. Stimuli, actions and rewards occur in the order of seconds and minutes, so that transient weights do hold their values during a learning phase. In effect, HTP suggests the intuitive notion that learning sessions may need to have a minimum duration or intensity of reward to be effective in the long term. Interestingly, experiments in human learning such as that described in Hamilton and Pascual-Leone (1998), seem to suggest that learning modifies synapses initially only in their short-term components, which decay within days if learning is suspended. A long lasting modification was registered only after months of training (Hamilton and Pascual-Leone, 1998). An intrigu-

ing possibility is that the consolidation of weights does not require months only because of biological limitations (e.g. growth of new synapses): the present model suggests that consolidation may require time in order to extract consistent and invariable relationships. So if short-term changes are consistently occurring across the same pathways every week for many weeks, long-term changes will also take place.

The model shows how neural structures may be preserved when learning. From this perspective, it emerges that the mechanism for learning is the same that preserves memory, effectively highlighting a strong coupling of learning and memory as it also suggested in biology (Bouton, 1994). It is nevertheless important to point out that the evidence of associative learning in animals (Grossberg, 1971; Bouton and Moody, 2004) depicts yet more complex dynamics that are not captured by current models.

Despite its simplified dynamics with respect to biological systems, the neural learning described by HTP offers a new tool to study learning and cognition both in animals and in neural artificial agents or neuro-robots (Krichmar and Roehrbein, 2013). The proposed dynamics allow for biological and robotics modelling of extended and realistic learning scenarios which were previously too complex for neural models. Examples are learning in interaction where overlapping stimuli, actions, and highly stochastic feedback occur at uncertain times (Soltoggio et al, 2013b). The acquisition of knowledge with HTP can integrate different tasks and scenarios, thereby opening the possibility of studying integrated cognition in unified neural models. This property may in turn result in models for the acquisition of incrementally complex behaviors at different stages of learning (Weng et al, 2001; Lungarella et al, 2003; Asada et al, 2009).

In the current model, long-term weights do not decay, i.e. they preserve their values indefinitely. This assumption reflects the fact that, if a certain relationship was established, i.e. if it was converted from hypothesis to certainty, it represents a fact in the world. To confirm that, the plot in Fig. 8a proved that, with a frequency of 1.5 Hz of the stimuli and a 100 ms sampling time, no wrong connection was consolidated in the extended experiment over 288 h of simulated time. The lack of reversal learning (long-term weights cannot decrease) works in this particular case because the environment and tasks in the current study are static, i.e. the stimulus-response pairs that induce rewards do not change. Under such conditions, the learning requires no unlearning. However, environments may be indeed changeable, and the rewarding conditions may change over time. In such cases, one simple extension for adaptation is nec-

essary. Assume that one rewarding pair ceases at one point to cause rewards. HTP will correctly detect the case by depressing the short-term weight, i.e. the hypothesis becomes negative. In the current algorithm, depression of short-term weights does not affect long-term weights. However, the consolidation described by Eq. (10) can be complemented by a symmetrical mechanism that depresses long-term weights when hypotheses are negative. With such an extension, the model can perform reversal of learning (Van Hemmen, 1997; Deco and Rolls, 2005; O'Doherty et al, 2001), thereby removing long-term connections when they do not represent anymore correct relationships in the world. The extension to unlearning is shown in the Appendix 1.

5 Conclusion

The proposed model introduces the concept of *hypothesis testing* of cause-effect relationships when learning with delayed rewards. The model describes a conceptual distinction between short-term and long-term plasticity, which is not focused on the duration of a memory, but it is rather related to the confidence with which cause-effect relationships are considered consistent (Abraham and Robins, 2005), and therefore preserved as memory.

The meta-plasticity rule, named *Hypothesis Testing Plasticity* (HTP), models how cause-effect relationships can be extracted from ambiguous information flows, first by validation and then by consolidation to long-term memory. The short-term dynamics boost exploration and discriminate more clearly true cause-effect relationships in a deceiving environment. The targeted conversion of short-term to long-term weights models the consolidation process of hypotheses in established facts, thereby addressing the plasticity-stability dilemma (Abraham and Robins, 2005). HTP suggests new cognitive models of biological and machine learning that explain dynamics of learning in complex and rich environments. This study proposes a theoretical motivation for short-term plasticity, which helps hypothesis testing, or learning in deceiving environments, and the following memorization and consolidation process.

Appendix 1: Unlearning

Unlearning of the long-term components of the weights can be effectively implemented as symmetrical to learning. I.e., when the transient weights are very negative (lower than $-\Psi$), the long-term component of a weight is decreased. This process represents the validation of the hypothesis that a certain stimulus-action pair is not associated with a reward anymore, or that is possibly associated with punishment. In such a case, the neural weight that represents this stimulus-action pair is decreased, and so is the probability of occurring. The conversion of negative transient weights to decrements of long-term weights, similarly to Eq. (10), can be formally expressed as

$$\dot{w}_{ji}^{lt}(t) = -\rho \cdot H(-w_{ji}^{st}(t) - \Psi) \quad . \tag{11}$$

No other changes are required to the algorithm described in the paper.

The case can be illustrated reproducing the preliminary test of Fig. 3, augmenting it with a phase characterised by a negative average modulation. Fig. 10 shows that, when modulatory updates become negative on average (from reward 4000 to reward 5000), the transient weight detects it by becoming negative. The use of Eq. (11) then causes the long-term component to reduce its value, thereby reversing the previous learning.

Preliminary experiments with unlearning on the complete neural model of this study show that the rate of negative modulation drops drastically as unlearning proceed. In other words, as the network experiences negative modulation, and consequently reduces the frequencies of punishing stimulus-action pairs, it also reduces the rate of unlearning because punishing episodes become sporadic. It appears that unlearning from negative experiences might be slower that learning from positive experiences. Evidence from biology indicates that extinction does not remove completely the previous association (Bouton, 2000, 2004), suggesting that more complex dynamics as those proposed here may regulate this process in animals.

Appendix 2: Implementation

All implementation details are also available as part of the open source Matlab code provided as support
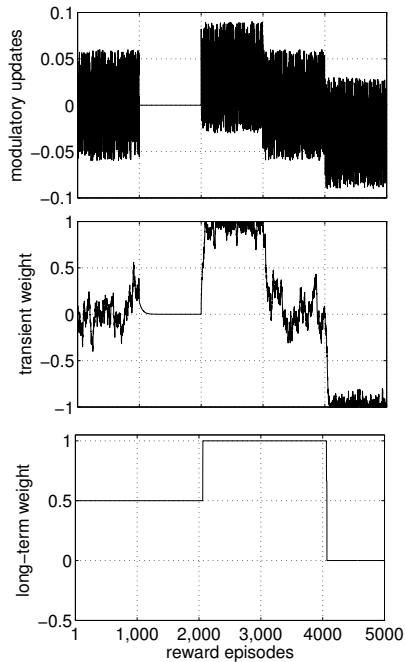
Fig. 10 Unlearning dynamics. In this experiment, the model presented in the paper was augmented with Eq. (11), which decreases long-term weights if the transient weights are lower than $-\Psi$. The stochastic modulatory update (top graph) is set to have a slightly negative average in the last phase (from reward 4001 to 5000). The negative average is detected by the short-term component that becomes negative. The long-term component decreases its value due to Eq. (11).

Parameter	Value
Inputs	300
Outputs	30
Stimulus/input duration	[1, 2] s
Max number of active inputs	3
Probability of no stimuli	1/8
Probability of 1 active stimulus	3/8
Probability of 2 active stimuli	3/8
Probability of 3 active stimuli	1/8
Action/output duration	[1, 2] s
Rewarding stimulus-action pairs	30
Delay of the reward	[1, 4] s
Nr of scenarios	3
Duration of one learning phase	24 h

Table 2 Summary of parameters for the input, output and reward signals.

material. The code can be used to reproduce the results in this work, or modified to perform further experiments. The source code can be downloaded from http://andrea.soltoggio.net/HTP.

Network, inputs, outputs, and rewards

The network is a feed-forward single layer neural network with 300 inputs, 30 outputs, 9000 weights, and sampling time of 0.1 s. Three hundred stimuli are delivered to the network by means of 300 input neurons. Thirty actions are performed by the network by means of 30 output neurons.

The flow of stimuli consists of a random sequence of stimuli each of duration between 1 and 2 s. The probability of 0, 1, 2 or 3 stimuli to be shown to the network simultaneously is described in Table 2.

The agent continuously performs actions chosen form a pool of 30 possibilities. Thirty output neurons may be interpreted as single neurons, or populations. When one action terminates, the output neuron with the highest activity initiates the next action. Once the response action is started, it lasts a variable time between 1 and 2 s. During this time, the neuron that initiated the action receives a feedback signal I of 0.5. The feedback current enables the output neuron responsible for one action to correlate correctly with the stimulus that is simultaneously active. A feedback signal is also used in Urbanczik and Senn (2009) to improve the reinforcement learning performance of a neural network.

The rewarding stimulus-action pairs are (i, i) with $1 \le i \le 10$ during scenario 1, $(i, i - 5)$ with $11 \le i \le 20$ in scenario 2, and $(i, i - 20)$ with $21 \le i \le 30$ in scenario 3. When a rewarding stimulus-action pair is performed, a reward is delivered to the network with a random delay in the interval [1, 4] s. Given the delay of the reward, and the frequency of stimuli and actions, a number of stimulus-action pairs could be responsible for triggering the reward. The parameters are listed in Table 2.

Integration

The integration of Eqs. (3) and (2) with a sampling time Δt of 100 ms is implemented step-wise by

$$E_{ji}(t + \Delta t) = E_{ji}(t) \cdot e^{\frac{-\Delta t}{\tau_E}} + \text{RCHP}_{ji}(t) \qquad (12)$$

$$m(t + \Delta t) = m(t) \cdot e^{\frac{-\Delta t}{\tau_m}} + \lambda\, r(t) + b \quad . \qquad (13)$$

The same integration method is used for all leaky integrators used in this study. Given that $r(t)$ is a signal from the environment, it might be a one-step signal as in the present study, which is high for one step when reward is delivered, or any other function representing a reward: in a test of RCHP on the real robot iCub (Soltoggio et al, 2013a,b), r(t) was determined by the human teacher by pressing skin sensors on the robots arms.

Parameter	Value
Number of neurons	330
Number of synapses	9000
Weight range	$[0, 1]$
Noise on neural transmission ($\xi_i(t)$, Eq. (6))	0.02 std
Sampling time step (Δt, Eq. (6))	100 ms
Baseline modulation (b in Eq. (2))	-0.03 / s
Neural gain (γ, Eq. (6))	0.5
Short-term learning rate (λ in Eqs. (2) and (13))	0.1
Time constant of modulation (τ_m)	0.1 s
Time constant of traces (τ_E)	4 s

Table 3 Summary of parameters of the neural model.

Parameter	Value
Rare correlations (μ in Eqs. (16) and (17))	$0.1\%/s$
Update rate of θ (η in Eqs. (16) and (17))	0.001 / s
α (Eq. (4))	1
β (Eq. (4))	1
Correlation sliding window (Eq. (15))	5 s
Short-term time constant (τ_{st} in Eq. (7))	8 h
Consolidation rate (ρ in Eq. (10))	1/1800 s
Consolidation threshold (Ψ in Eq. (10))	0.95

Table 4 Summary of parameters of the plasticity rules (RCHP and RCHP$^+$ plus HTP).

Rarely Correlating Hebbian Plasticity

Rarely Correlating Hebbian Plasticity (RCHP) (Soltoggio and Steil, 2013) is a type of Hebbian plasticity that filters out the majority of correlations and produces nonzero values only for a small percentage of synapses. Rate-based neurons can use a Hebbian rule augmented with two thresholds to extract low percentages of correlations and decorrelations. RCHP expressed by Eq. (4) is simulated with the parameters in Table 4. The rate of correlations can be expressed by a global concentration ω_c. This measure represents how much the activity of the network correlates, i.e. how much the network activity is deterministically driven by connections or is instead noise-driven. The instantaneous matrix of correlations RCHP$^+$ (i.e. the first row in Eq. (4) computed for all synapses) can be low filtered as

$$\dot{\omega}_c(t) = -\frac{\omega_c(t)}{\tau_c} + \sum_{j=1}^{300}\sum_{i=1}^{30} \text{RCHP}^+_{ji}(t) \quad , \tag{14}$$

to estimate the level of correlations in the recent past, where j is the index of input neurons, and i the index of the output neurons. In the current settings, τ_c was chosen equal to 5 s. Alternatively, a similar measure of recent correlations $\omega_c(t)$ can be computed in discrete time over a sliding time window of 5 s summing all correlations RCHP$^+(t)$

$$\omega_c(t) = \Delta t \frac{\sum_0^{t-5} \text{RCHP}^+(t)}{5} \quad . \tag{15}$$

Similar equations to (14) and (15) are used to estimate decorrelations $\omega_d(t)$ from the detected decorrelations RCHP$^-(t)$. The adaptive thresholds θ_{hi} and θ_{lo} in Eq. (4) are estimated as follows.

$$\theta_{hi}(t + \Delta t) = \begin{cases} \theta_{hi} + \eta \cdot \Delta t & \text{if } \omega_c(t) > 2\mu \\ \theta_{hi} - \eta \cdot \Delta t & \text{if } \omega_c(t) < \mu/2 \\ \theta_{hi}(t) & \text{otherwise} \end{cases} \tag{16}$$

and

$$\theta_{lo}(t + \Delta t) = \begin{cases} \theta_{lo} - \eta \cdot \Delta t & \text{if } \omega_d(t) > 2\mu \\ \theta_{lo} + \eta \cdot \Delta t & \text{if } \omega_d(t) < \mu/2 \\ \theta_{lo}(t) & \text{otherwise} \end{cases} \tag{17}$$

with $\eta = 0.001$ and μ, the target rate of rare correlations, set to 0.1%/s. If correlations are lower than half of the target or are greater than twice the target, the thresholds are adapted to the new increased or reduced activity. This heuristic has the purpose of maintaining the thresholds relatively constant and perform adaptation only when correlations are too high or too low for a long period of time.

References

Abbott LF, Regehr WG (2004) Synaptic computation. Nature 431:796–803

Abraham WC (2008) Metaplasticity: tuning synapses and networks for plasticity. Nature Reviews Neuroscience 9:387–399

Abraham WC, Bear MF (1996) Metaplasticity: the plasticity of synaptic plasticity. Trends in Neuroscience 19:126–130

Abraham WC, Robins A (2005) Memory retention–the synaptic stability versus plasticity dilemma. Trends in Neuroscience 28:73–78

Alexander WH, Sporns O (2002) An Embodied Model of Learning, Plasticity, and Reward. Adaptive Behavior 10(3-4):143–159

Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: a survey. Autonomous Mental Development, IEEE Transactions on 1(1):12–34

Bailey CH, Giustetto M, Huang YY, Hawkins RD, Kandel ER (2000) Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? Nature Reviews Neuroscience 1(1):11–20

Baras D, Meir R (2007) Reinforcement Learning, Spike-Time-Dependent plasticity, and the BCM Rule. Neural Computation 19(8):2245–2279

Ben-Gal I (2007) Bayesian Networks, in: Encyclopedia of Statistics in Quality and Reliability, Wiley & Sons

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191:391–431

Bosman R, van Leeuwen W, Wemmenhove B (2004) Combining Hebbian and reinforcement learning in a minibrain model. Neural Networks 17:29–36

Bouton ME (1994) Conditioning, remembering, and forgetting. Journal of Experimental Psychology: Animal Behavior Processes 20(3):219

Bouton ME (2000) A learning theory perspective on lapse, relapse, and the maintenance of behavior change. Health Psychology 19(1S):57

Bouton ME (2004) Context and behavioral processes in extinction. Learning & memory 11(5):485–494

Bouton ME, Moody EW (2004) Memory processes in classical conditioning. Neuroscience & Biobehavioral Reviews 28(7):663–674

Brembs B (2003) Operant conditioning in invertebrates. Current opinion in neurobiology 13(6):710–717

Brembs B, Lorenzetti FD, Reyes FD, Baxter DA, Byrne JH (2002) Operant Reward Learning in Aplysia: Neuronal Correlates and Mechanisms. Science 296(5573):1706–1709

Bullinaria JA (2009) Evolved dual weight neural architectures to facilitate incremental learning. In: Proceedings of the International Joint Conference on Computational Intelligence (IJCCI 2009), pp 427–434

Clopath C, Ziegler L, Vasilaki E, Büsing L, Gerstner W (2008) Tag-trigger-consolidation: A model of early and late long-term-potentiation and depression. PLoS Computational Biology 4(12):335.347

Cox RB, Krichmar JL (2009) Neuromodulation as a robot controller: A brain inspired strategy for controlling autonomous robots. IEEE Robotics & Automation Magazine 16(3):72–80

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

Dudai Y (2004) The neurobiology of consolidations, or, how stable is the engram? Annual Review of Psychology 55:51–86

Farries MA, Fairhall AL (2007) Reinforcement Learning With Modulated Spike Timing-Dependent Synaptic Plasticity. Journal of Neurophysiology 98:3648–3665

Fisher SA, Fischer TM, Carew TJ (1997) Multiple overlapping processes underlying short-term synaptic enhancement. Trends in neurosciences 20(4):170–177

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

Frémaux N, Sprekeler H, Gerstner W (2010) Functional requirements for reward-modulated spike-timing-dependent plasticity. The Journal of Neuroscience 30(40):13,326–13,337

Frey U, Morris RGM (1997) Synaptic tagging and long-term potentiation. Nature 385(533-536)

Friedrich J, Urbanczik R, Senn W (2010) Learning spike-based population codes by reward and population feedback. Neural Computation 22:1698–1717

Friedrich J, Urbanczik R, Senn W (2011) Spatio-temporal credit assignment in neuronal population learning. PLoS Comput Biol 7(6):1–13

Fusi S, Senn W (2006) Eluding oblivion with smart stochastic selection of synaptic updates. Chaos: An Interdisciplinary Journal of Nonlinear Science 16(2):026,112

Fusi S, Drew PJ, Abbott L (2005) Cascade models of synaptically stored memories. Neuron 45(4):599–611

Fusi S, Asaad WF, Miller EK, Wang XJ (2007) A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron 54(2):319–333

Garris P, Ciolkowski E, Pastore P, 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 (2010) From Hebb rules to spike-timing-dependent plasticity: a personal account. Frontiers in Synaptic Neuroscience 2:1–3

Gil M, DeMarco RJ, Menzel R (2007) Learning reward expectations in honeybees. Learning and Memory 14:291–496

Goelet P, Castellucci VF, Schacher S, Kandel ER (1986) The long and the short of long-term memory: A molecular framework. Nature 322:419–422

Grossberg S (1971) On the dynamics of operant conditioning. Journal of Theoretical Biology 33(2):225–255

Grossberg S (1988) Nonlinear neural networks: principles, mechanisms, and architectures. Neural Networks 1:17–61

Hamilton RH, Pascual-Leone A (1998) Cortical plasticity associated with braille learning. Trends in cognitive sciences 2(5):168–174

Hammer M, Menzel R (1995) Learning and memory in the honeybee. The Journal of Neuroscience 15(3):1617–1630

Heckerman D, Geiger D, Chickering DM (1995) Learning bayesian networks: The combination of knowledge and statistical data. Machine Learning 20:197–243

Hinton GE, Plaut DC (1987) Using fast weights to deblur old memories. In: Proceedings of the ninth annual conference of the Cognitive Science Society, Erlbaum, pp 177–186

Howson C, Urbach P (1989) Scientific reasoning: The Bayesian approach. Open Court Publishing Co, Chicago, USA

Hull CL (1943) Principles of behavior. New-Your: Appleton Century

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

Jay MT (2003) Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. Progress in Neurobiology 69(6):375–390

Jonides J, Lewis RL, Nee DE, Lustig CA, Berman MG, Moore KS (2008) The mind and brain of short-term memory. Annual review of psychology 59:193

Kempter R, Gerstner W, Van Hemmen JL (1999) Hebbian learning and spiking neurons. Physical Review E 59(4):4498–4514

Krichmar JL, Roehrbein F (2013) Value and reward based learning in neurorobots. Frontiers in Neurorobotics 7(13)

Lamprecht R, LeDoux J (2004) Structural plasticity and memory. Nature Reviews Neuroscience 5(1):45–54

Legenstein R, Chase SM, Schwartz A, Maass W (2010) A Reward-Modulated Hebbian Learning Rule Can Explain Experimentally Observed Network Reorganization in a Brain Control Task. The Journal of Neuroscience 30(25):8400–8401

Leibold C, Kempter R (2008) Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. Cerebral Cortex 18(1):67–77

Levy JP, Bairaktaris D (1995) Connectionist dual-weight architectures. Language and Cognitive Processes 10(3-4):265–283

Lin LJ (1993) Reinforcement learning for robots using neural networks. PhD thesis, School of Computer Science, Carnegie Mellon University

Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connection Science 15(4):151–190

Lynch MA (2004) Long-term potentiation and memory. Physiological Reviews 84(1):87–136

Mayford M, Siegelbaum SA, Kandel ER (2012) Synapses and memory storage. Cold Spring Harbor perspectives in biology 4(6):a005,751

McGaugh JL (2000) Memory–a century of consolidation. Science 287:248–251

Menzel R, Müller U (1996) Learning and Memory in Honeybees: From Behavior to Natural Substrates. Annual Review of Neuroscience 19:179–404

Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive Hebbian learning. Nature 377:725–728

Nguyen PV, Abel T, Kandel ER (1994) Requirement of a critical period of transcription for induction of a late phase of ltp. Science 265(5175):1104–1107

Nitz DA, Kargo WJ, Fleisher J (2007) Dopamine signaling and the distal reward problem. Learning and Memory 18(17):1833–1836

O'Brien MJ, Srinivasan N (2013) A Spiking Neural Model for Stable Reinforcement of Synapses Based on Multiple Distal Rewards. Neural Computation 25(1):123–156

O'Doherty JP, Kringelbach ML, Rolls ET, Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. Nature Neuroscience 4(1):95–102

Ono K (1987) Superstitious behavior in humans. Journal of the Experimental Analysis of Behavior 47(3):261–271

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

Pennartz CMA (1996) The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. Brain Research Reviews 21:219–245

Pennartz CMA (1997) Reinforcement Learning by Hebbian Synapses with Adaptive Threshold. Neuroscience 81(2):303–319

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

Robins A (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. Connection Science: Journal of Neural Computing, Artificial Intelligence and Cognitive Research 7(123-146)

Sandberg A, Tegnér J, Lansner A (2003) A working memory model based on fast hebbian learning. Network: Computation in Neural Systems 14(4):789–802

Sarkisov DV, Wang SSH (2008) Order-Dependent Coincidence Detection in Cerebellar Purkinje Neurons at the Inositol Trisphosphate Receptor. The Journal of Neuroscience 28(1):133–142

Schmidhuber J (1992) Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. Neural Computation 4:131–139

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

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

Senn W, Fusi S (2005) Learning only when necessary: better memories of correlated patterns in networks with bounded synapses. Neural Computation 17(10):2106–2138

Skinner BF (1948) "Superstition" in the pigeon. Journal of Experimental Psychology 38:168–172

Skinner BF (1953) Science and Human Behavior. New York, MacMillan

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

Soltoggio A, Steil JJ (2013) Solving the Distal Reward Problem with Rare Correlations. Neural Computation 25(4):940–978

Soltoggio A, Bullinaria JA, Mattiussi C, Dürr P, 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, Lemme A, Reinhart FR, Steil JJ (2013a) Rare neural correlations implement robotic conditioning with reward delays and disturbances. Frontiers in Neurorobotics 7(Research Topic: Value and Reward Based Learning in Neurobots)

Soltoggio A, Reinhart FR, Lemme A, Steil JJ (2013b) Learning the rules of a game: neural conditioning in human-robot interaction with delayed rewards. In: Proceedings of the Third Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics - Osaka, Japan - August 2013

Sporns O, Alexander WH (2002) Neuromodulation and plasticity in an autonomous robot. Neural Networks 15:761–774

Sporns O, Alexander WH (2003) Neuromodulation in a learning robot: interactions between neural plasticity and behavior. In: Proceedings of the International Joint Conference on Neural Networks, vol 4, pp 2789–2794

Staubli U, Fraser D, Faraday R, Lynch G (1987) Olfaction and the "data" memory system in rats. Behavioral Neuroscience 101(6):757–765

Sutton RS (1984) Temporal credit assignment in reinforcement learning. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA, USA

Swartzentruber D (1995) Modulatory mechanisms in pavlovian conditioning. Animal Learning & Behavior 23(2):123–143

Thorndike EL (1911) Animal Intelligence. New York: Macmillan

Tieleman T, Hinton G (2009) Using fast weights to improve persistent contrastive divergence. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, pp 1033–1040

Urbanczik R, Senn W (2009) Reinforcement learning in populations of spiking neurons. Nature Neuroscience 12:250–252

Van Hemmen J (1997) Hebbian learning, its correlation catastrophe, and unlearning. Network: Computation in Neural Systems 8(3):V1–V17

Wang SSH, Denk W, Häusser M (2000) Coincidence detection in single dendritic spines mediated by calcium release. Nature Neuroscience 3(12):1266–1273

Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E (2001) Autonomous mental development by robots and animals. Science 291(5504):599–600

Wighmann R, 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 RA, Rompre PP (1989) Brain dopamine and reward. Annual Review of Psychology 40:191–225

Xie X, Seung HS (2004) Learning in neural networks by reinforcement of irregular spiking. Physical Review E 69:1–10

Ziemke T, Thieme M (2002) Neuromodulation of Reactive Sensorimotor Mappings as Short-Term Memory Mechanism in Delayed Response Tasks. Adaptive Behavior 10:185–199

Zucker RS (1989) Short-term synaptic plasticity. Annual review of neuroscience 12(1):13–31

Zucker RS, Regehr WG (2002) Short-term synaptic plasticity. Annual review of physiology 64(1):355–405