From modulated Hebbian plasticity to simple behavior learning through noise and weight saturation

From Modulated Hebbian Plasticity to Simple Behavior Learning through Noise and Weight Saturation

Preprint accepted for publication in Neural Networks (2012) doi: 10.1016/j.neunet.2012.06.005

Andrea Soltoggio[a], Kenneth O. Stanley[b]

[a] *Research Institute for Cognition and Robotics, University of Bielefeld, Germany, asoltogg@cor-lab.uni-bi.de*
[b] *Dept. of Electrical Engineering and Computer Science, University of Central Florida, USA, kstanley@eecs.ucf.edu*

Abstract

Synaptic plasticity is a major mechanism for adaptation, learning and memory. Yet current models struggle to link local synaptic changes to the acquisition of behaviors. The aim of this paper is to demonstrate a computational relationship between local Hebbian plasticity and behavior learning by exploiting two traditionally unwanted features: neural noise and synaptic weight saturation. A modulation signal is employed to arbitrate the sign of plasticity: when the modulation is positive, the synaptic weights saturate to express exploitative behavior; when it is negative, the weights converge to average values and neural noise reconfigures the network's functionality. This process is demonstrated through simulating neural dynamics in the autonomous emergence of fearful and aggressive navigating behaviors and in the solution to reward-based problems. The neural model learns, memorizes and modifies different behaviors that lead to positive modulation in a variety of settings. The algorithm establishes a simple relationship between local plasticity and behavior learning by demonstrating the utility of noise and weight saturation. Moreover it provides a new tool to simulate adaptive behavior and contributes to bridging the gap between synaptic changes and behavior in neural computation.

Keywords:
Adaptive behavior, Computational models, Learning, Neural plasticity, Neuromodulation

1. Introduction

This paper describes a novel, modulated Hebbian plasticity rule that makes productive use of features of Hebbian dynamics that in the past were thought undesirable. By utilizing noise and saturation, operant reward learning can emerge from the present learning rule, establishing an important link between local plasticity and macro-level behavioral adaptation.

The idea that adaptation, learning and memory rely on synaptic change has gathered increasing consensus beginning with the early studies of Hebb (1949) and the seminal work of Kandel and Tauc (1965). Early studies on the mollusk Aplysia proved that behavioral changes were precisely linked to the growth of particular pathways from sensory to motor systems (Kandel and Tauc, 1965; Carew et al., 1981). However, synaptic change follows rich dynamics that are often the product of different chemical signals (Clark, 2001) whose interaction and mechanisms are not completely understood. The Hebbian paradigm (Hebb, 1949; Marr, 1969; Stent, 1973; Brown et al., 1990; Bi and Poo, 2001; Gerstner and Kistler, 2002a; Cooper, 2005), which states that *neurons that fire together, wire together*, is a ubiquitous paradigm in neuroscience that has been substantially validated through neural recordings (Stent, 1973; Kelso et al., 1986; McNaughton et al., 1986; Lisman, 1989; Markram et al., 1997), corroborating detailed rate-based (Grossberg, 1976; Rauschecker and Singer, 1981; Oja, 1982; Bienenstock et al., 1982; Gerstner and Kistler, 2002a) and spiking neural models (van Rossum et al., 2000).

The increasingly evident link between behavior learning and synaptic plasticity has encouraged researchers to propose numerous models whose overall behavior changes with the modification of synaptic weights; for reviews see Dayan and Abbott (2001); Bi and Poo (2001); Gerstner and Kistler (2002a). However, one controversial and often unwanted feature of Hebbian models is that increasing firing leads to increasing synaptic strength, which in turn leads to further increasing of firing (Miller and Mackay, 1994; Hasselmo, 1994; Moldakarimov and Sejnowski, 2008). When a weight is larger than a certain threshold, a positive feedback will cause the weight to increase indefinitely. Such a model yields autocorrelation rather than cross-correlation of signals (Porr and Wörgötter, 2006). To prevent indefinite weight growth, various constraints can be imposed on the basic Hebbian plasticity (Oja, 1982; Bienenstock et al., 1982; Miller and Mackay, 1994). A second limitation of simple Hebbian plasticity is that learning can be just as fast

as unlearning. For such models, short-lived stimuli leave a short-lived trace in the network regardless of their relevance. This feature contrasts with long term potentiation (LTP), in which certain conditions induce synapses to maintain the increased strength in the long term (Levy and Steward, 1979, 1983; Kelso et al., 1986; Brown et al., 1988; Gustafsson et al., 1987).

In effect, the dynamics of Hebbian plasticity in biology are often affected and substantially altered by additional homeostatic dynamics (Turrigiano, 2008) and neuromodulators (Harris-Warrick and Marder, 1991; Hasselmo, 1995; Giocomo and Hasselmo, 2007; Bailey et al., 2000; Clark, 2001). For example, when the Aplysia encounters noxious stimuli, additional modulatory activity is also triggered, resulting in longer-lasting synaptic changes (Clark and Kandel, 1984; Bailey et al., 2000). This observation suggests that additional modulatory chemicals act as selectors of relevant stimuli that require learning of long-lasting responses, as in the case of dangerous or pain-inducing conditions (Bailey et al., 2000). To date there is extensive evidence linking conditioning behavior and reward learning with neuromodulation. Modulatory activity appears to carry reward information across a surprisingly large spectrum of animals, from insects like the honeybee (Hammer, 1993; Gil et al., 2007), to mollusks like the Aplysia (Walters and Byrne, 1983; Brembs et al., 2002) and to mammals (Schultz et al., 1993, 1997; Wise and Rompre, 1989; Berridge and Robinson, 1998). Yet whether and why neuromodulation is computationally essential to achieve such long-lasting behavioral responses has not been clarified.

Driven by biological findings, researchers have augmented their models with modulatory signals (Hasselmo and Schnell, 1994; Fellous and Linster, 1998; Ludvig et al., 2008) or attempted to model biological modulatory activities (Baxter et al., 1999; Cohen, 2008). The precise role of various modulatory chemicals (e.g. serotonin, acetylcholine, dopamine and norepinephrine (Bear et al., 2005; Hasselmo, 2006)) is still debated, in particular regarding the role of dopamine in reward learning (Pennartz, 1996, 1997; Berridge and Robinson, 1998; Montague et al., 2004; Schultz, 2006; Redgrave et al., 2008). Moreover, modulation appears to regulate a large variety of behaviors like arousal, attention, reward learning, and memory (Harris-Warrick and Marder, 1991; Hasselmo, 1995; Aston-Jones and Cohen, 2005), resulting in an accordingly large spectrum of dynamics and models that regulate synaptic efficacy, synaptic changes and other neural variables (Hasselmo and Schnell, 1994; Fellous and Linster, 1998; Doya, 2002; Smith et al., 2002; Krichmar, 2008; Cox and Krichmar, 2009). One promising computational aspect of modulation is the possibility of increasing, decreasing or inverting the strength and sign of plasticity (Abbott, 1990; Montague et al., 1996; Florian, 2007; Porr and Wörgötter, 2007; Izhikevich, 2007; Pfeiffer et al., 2010), making neuromodulation particularly suitable for modeling and implementing learning processes (Sporns and Alexander, 2002; Doya, 2002; Doya and Uchibe, 2005; Soula et al., 2005; Farries and Fairhall, 2007; Krichmar, 2008; Cox and Krichmar, 2009). The focus in this study is on this latter role of modulation as a gating mechanism for Hebbian synaptic plasticity.

A fundamental issue is that a weight change that follows local rules does not always have a straightforward relationship with the system-level input-output mapping. This disconnect makes it difficult to apply local unsupervised plasticity rules to the fields of simulated adaptive behavior, artificial life (Langton, 1990; Sporns and Alexander, 2002) and robotics (Arkin, 1998). In these areas, the use of closed-loop controllers, in which the relationships between local and system-level dynamics are continuously tested, can provide the ultimate verification of the learning properties of a model. The model presented in this paper aims to establish a simple relationship between modulated Hebbian plasticity and operant reward learning, thereby connecting models of plasticity more closely to the learning of behaviors.

Instead of focusing on precise weight tuning, the unique position of this paper is to search for behavioral responses by allowing the weights to saturate, expressing either highly excitatory or inhibitory responses. By intentionally allowing weights to saturate, a network can express a marked and stable response to inputs, which can be interpreted as behavioral *exploitation*. On the other hand, by inverting this process at times, i.e. by inverting the sign of Hebbian plasticity (Stent, 1973; Lisman, 1989), pathways can be depressed to allow noisy neural transmission to implement behavioral *exploration*. The alternation of these two regimes of Hebbian and anti-Hebbian plasticity produces the key dynamics of alternating exploitation and exploration observed in operant reward learning. The change in modulatory activity has in fact been suggested to regulate the alternation of exploration and exploitation in Krichmar (2008). Thus, while the dynamics of modulated Hebbian plasticity and modulated spike-time-dependent plasticity (STDP) have been extensively investigated (Abbott, 1990; Montague et al., 1996; Florian, 2007; Porr and Wörgötter, 2007; Frémaux et al., 2010; Pfeiffer et al., 2010), the novelty of this work is their extension by means of saturation and noise, resulting in a simpler and more fundamental connection between local changes and higher-level simulated behavior. The fundamental properties of the new plasticity model are tested in behavioral tasks employing first a single-neuron model, and later extended to multi-neuron networks.

As opposed to the algorithms proposed by Pfeiffer et al. (2010), Legenstein et al. (2010) and Frémaux et al. (2010), the present work neither devises a learning rule for optimal weight tuning nor proposes a new reinforcement learning algorithm. In fact, while reinforcement learning by means of modulated spike-timing dependent plasticity (STDP) was demonstrated in Soula et al. (2005), Florian (2007) and Frémaux et al. (2010), the primary aim of this work is the exploitation of saturated weights and neural noise to achieve a simple bottom-up implementation of oper-

ant reward learning. Furthermore, in contrast to Pfeiffer et al. (2010), the current algorithm does not require a decay function, input signal preprocessing nor winner-take-all action selection. Crucially, the neural noise in the present implementation is not used to *improve* exploration, as in Legenstein et al. (2010), but rather serves as the only and fundamental driving mechanism to reconfigure the network connectivity, thereby achieving behavioral exploration under the anti-Hebbian regime. Additionally, as opposed to Legenstein et al. (2010), where slow variation of input values and continuity in the task are required for recent activity averages, inputs and outputs in the proposed method can change state arbitrarily according to sudden changes of the task or environmental conditions. The insight that operant reward learning can emerge naturally and without additional engineering from Hebbian dynamics is a fundamental contribution of this study.

The plasticity mechanism, described in Section 2, is tested in several simulations reported in Section 3. Section 4 discusses the results, and Section 5 presents the conclusion. Appendix A reports that the plasticity rule behaves similarly on a simple spiking-neuron model. Further implementation details and how to reproduce the results with the Matlab code are reported in Appendices B and C.

2. Reconfigure-and-saturate Hebbian plasticity

In a rate-based model, the simplest form of Hebbian plasticity is expressed by the product of presynaptic and postsynaptic firing rates

$$\Delta w_{ji} = v_j \cdot v_i \,, \tag{1}$$

where w_{ji} is the weight from neuron j to neuron i and the firing rate v is computed as a nonlinear monotonically increasing function, e.g. the hyperbolic tangent, of the membrane potential. The membrane potential is the weighted sum of the incoming firing rates.

When a presynaptic signal increases the activation of a postsynaptic neuron, Eq. 1 causes the weight to increase, which in turn causes a stronger correlation of activities. The auto-correlative dynamics then lead to indefinite weight growth (Miller and Mackay, 1994). The unique position in this paper is that this auto-correlative effect can be beneficial. The dynamics induced by Eq. 1 in fact implement weight consolidation (Fusi et al., 2000; Gerstner and Kistler, 2002a) by enhancing pathways among correlating neurons and suppressing pathways between non-correlating neurons. As suggested by Fusi et al. (2000), dynamics that favor either maximum or minimum weights can help preserve information. In the current model, indefinite weight growth is counteracted only by hard bounds between zero and a saturation value λ.

Consider now one excitatory and one inhibitory weight carrying the same signal and projecting onto the same postsynaptic neuron. In the model of this study, excitatory and inhibitory neurons are distinct entities, and are labeled

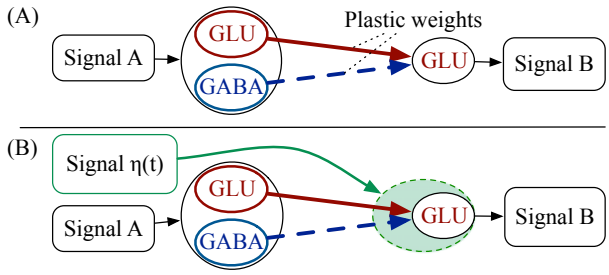

Figure 1: Pathways of two competing weights. (A) Signal A represents an input or an internal signal to be transformed into Signal B. Signal A feeds a cluster of two excitatory (GLU, from glutamate) and inhibitory (GABA, from GABAergic; see (Bear et al., 2005)) neurons, which then project onto a postsynaptic neuron that encodes Signal B. The circles can be interpreted either as a single neuron or a cluster of neurons of the same type. Given an initial random configuration, one of the two weights may be stronger than the other. The stronger weight drives the activity of the postsynaptic neuron. When the hyperbolic tangent produces the output, or in any other case when tonic activity produces a null output and inhibited activity produces a negative output, the effect of Eq. 1 is to increase the stronger weight and reduce the weaker. (B) The addition of a modulatory signal allows the structure to reverse the divergence of weights by reducing them to similar values.

glutamate (GLU) neurons and GABAergic (GABA) neurons after the corresponding neurotransmitters in biology (Bear et al., 2005). Fig. 1*A* illustrates the scenario. This structure of two competing weights carrying the same signal is called in this study a *pathway*. If the two weights have equal strength, the output neuron is neither excited nor inhibited. If one of the two weights is stronger, the activation of the output neuron follows the stronger weight. As observed by Pennartz (1997), models in which synaptic connections change sign are not biologically implausible and can be interpreted as two such parallel excitatory and inhibitory weights (Alger and Nicoll, 1982; Finch et al., 1988; Pennartz and Kitai, 1991). In the proposed model, when the GABA input prevails over the GLU input and successfully inhibits the output, the GABA-to-output weight increases its strength (Stent, 1973). This means that in this model inhibition is not modeled as a homeostatic signal, i.e. it does not bring the network to an equilibrium because inhibition causes even more inhibition. However, this process can be reversed by changing the sign of Hebbian plasticity by means of a modulatory term η (Abbott, 1990; Porr and Wörgötter, 2007; Florian, 2007; Soltoggio et al., 2008; Pfeiffer et al., 2010) in

$$\Delta w_{ji} = \eta \cdot v_j \cdot v_i \,. \tag{2}$$

The term η (Fig. 1*B*) can be interpreted as a third factor, or *modulatory signal*. A positive unitary η is the simplest form of Hebbian plasticity. A negative unitary η is the simplest anti-Hebbian form as described in Brown et al. (1990). The signal η can be computed by internal neural structures as proposed by Soltoggio et al. (2007, 2008), or, for simple problems, it can be directly derived from a sensory input (Montague et al., 1996). In both cases,

η is derived directly from, or is in relation to, the agent-environment interaction, and its value is problem and contingency dependent.

The modulatory signal η acts as a referee determining whether to increase or decrease in absolute value the overall strength of a pathway. Positive modulation increases the difference between the two weights and negative modulation decreases the difference, thereby decreasing the overall strength of the pathway. A low modulatory value, or tonic value, implies small changes and is useful in conditions of low relevance for learning, as opposed to phasic values (Aston-Jones and Cohen, 2005; Krichmar, 2008). A crucial aspect in the present implementation is that under negative modulation the pathway oscillates between weakly excitatory and inhibitory states due to the stabilizing effect of anti-Hebbian plasticity. In this condition, neural noise becomes determinant in driving the neural dynamics to random alternations of the pathway's sign. These random alternations are the key element in the exploration of new neural configurations when a network includes such pathways. These dynamics of alternating random reconfigurations and then growing the weights to saturation are the core of the proposed model. This rule is therefore called *reconfigure-and-saturate* modulated Hebbian plasticity.

The neural output signals are then computed as

$$v_i(t+1) = \tanh\Big(\sum_{j=1}^{i=n} w_{ji}(t) \cdot v_j(t)\Big) + \xi_i(t)\,, \qquad (3)$$

where w_{ji} is the weight from neuron j to neuron i, and $\xi_i(t)$ is a random number drawn from a uniform distribution in [-0.1,0.1] unless differently specified. Note that the neuron inputs at time t (i.e. $v_j(t)$) produce the output at time $t+1$. This time delay can be interpreted as the propagation time of a signal between presynaptic and postsynaptic neurons, or alternatively as the single neuron computation delay. The weight update in the reconfigure-and-saturate plasticity rule is expressed by

$$w^*_{ji}(t) = w_{ji}(t-1) + \Big(C \cdot \eta(t) \cdot v_i(t) \cdot v_j(t-1)\Big) + \xi_{ji}(t) \qquad (4)$$

$$w_{ji}(t) = \max(0, \min(w^*_{ji}(t), \lambda))\,, \qquad (5)$$

which is applied on two excitatory and inhibitory competing weights carrying the same signal. The variable C is +1 when the presynaptic neuron is excitatory and -1 when the presynaptic neuron is inhibitory, v_j and v_i are the outputs of the presynaptic and postsynaptic neuron, respectively, and η is the modulatory signal. Eq. 5 maintains the weight values in the interval $[0,\ \lambda]$. The factor C indicates that when a presynaptic inhibitory neuron successfully inhibits a postsynaptic neuron, these are seen as correlating activities. Note that the presynaptic activity is taken at the time $t-1$ because, according to Eq. 3, this signal affects the postsynaptic neuron only at the following time step. Therefore, Eq. 4 expresses causality of firing rather than simultaneity. Causality is a governing principle in spike-timing-dependent plasticity (STDP) (Markram et al., 1997). It is nevertheless important to note that the dynamics of the reconfigure-and-saturate rule are considerably simplified with respect to their inspiring biological counterparts; the current study does not aim to mimic biological neural activity.

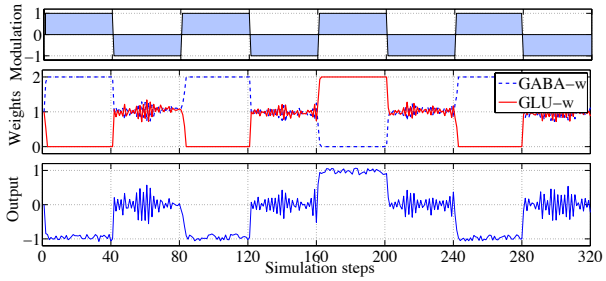

Figure 2: Simulation of the reconfigure-and-saturate modulated Hebbian plasticity on a one-input one-output structure. The first row shows the imposed modulation, whose intensity (either positive or negative) is visually represented by the colored areas. The second row shows the GLU (excitatory) and GABA (inhibitory) weight values. Finally, the third row shows the output signal. The output is neutral during negative modulation, whereas it assumes either positive or negative values when the modulation is positive. The sign of the output during positive modulation is a stochastic process depending on the initial conditions and neural noise.

As opposed to previous studies, the novelty introduced by this new rule is the fundamental exploratory role of noise level $\xi(t)$ during anti-Hebbian phases in combination with the convergence to saturation values λ during Hebbian phases. How these two factors contribute exactly to the dynamics of the reconfigure-and-saturate plasticity rule is investigated in the following sections.

3. Experiments

The simulations presented in this section explore the effect of the reconfigure-and-saturate Hebbian plasticity in increasingly complex scenarios. First, the effect of modulation change is observed on a single pathway of two competing weights. Next, the behavioral consequences of the weight dynamics are demonstrated in a navigating agent. Finally, the model is tested on problems with an arbitrary number of inputs and outputs. All experiments and figures with output data can be easily reproduced running the Matlab® code provided at this article's associated website `http://andrea.soltoggio.net/rec-sat`.

3.1. Effect of modulation change on a single pathway

The purpose of this test is to observe the effect of changing the modulation sign on the mapping from input to output. Therefore, while the input (Signal A in Fig. 1*B*) is held constantly high, the modulation signal (η) alternates from +1 to -1 every 40 steps.

Fig. 2 shows that the GLU and GABA weights diverge when the modulation is positive, causing one of the two

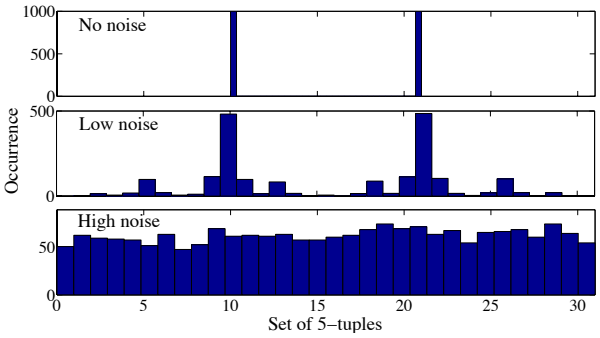

Figure 3: Testing exploration in the output sequence. The histograms count the occurrence of each of the 32 different tuples of 5 bits found in the output sequence. The top histogram (the run with no noise) shows that only the tuples 01010 and 10101 (10 and 21 in decimal notation) are represented. The middle histogram (the run with low noise) shows that other tuples begin to appear, but the 01010 and 10101 tuples are predominant. The bottom histogram (the run with high noise) shows that all possible tuples are similarly represented. This result means that with high noise, each reconfiguration leads the network to a new random state.

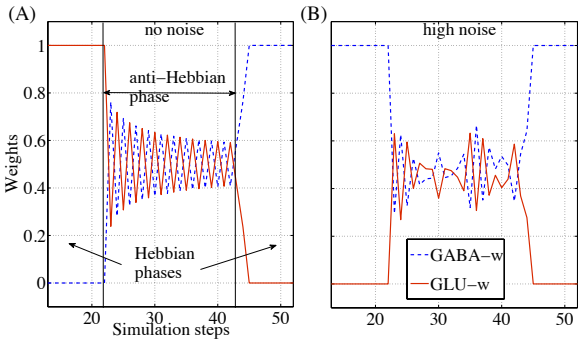

Figure 4: (A) Weights during an anti-Hebbian phase between two Hebbian phases. In the absence of noise, the stronger and weaker weights have regular oscillations. (B) With noise, the oscillations are perturbed, resulting in a less predictable alternation of weights. The addition of noise does not affect the weight values during positive modulation.

weights to prevail over the other. When the modulation becomes negative, the two weights converge and oscillate around the same value, neither of them prevailing. During this phase the input signal is not converted into a definite output value because the latter oscillates around zero. When the modulation returns to being positive, whichever of the two weights is higher during the oscillation at that moment prevails over the other, leading the output either to continuous excitation or continuous inhibition. This property is retained when simulating the model with spiking neurons, as shown in Appendix A.

3.2. The role of noise

This section shows that neural noise is essential to exploring *all network states* in networks of one neuron to networks of many. The previous experiment was performed again with three different settings: no noise, low noise ($\epsilon = 0.02$) and high noise ($\epsilon = 0.1$), corresponding to 0, 2% and 10%, respectively, of the maximum output value. The output is sampled at the end of each *positive* modulation period, i.e. when the weights and the output are stable at the saturation level, e.g. in Fig. 2 at steps 40, 120, 200, etc. This condition of saturated weights is called in the current study a *network configuration state*. The aim is to observe the sequence of configuration states over a long simulation. For each noise level, 10,000 modulation periods were executed, producing a binary string that represents the sequence of configuration states. The network reconfigures to a random state at each modulation phase if the observed sequence follows a binomial distribution.

A simple stochasticity test was conducted by segmenting the sequence into 5-tuples and counting the occurrences of each possible 5-tuple of binary digits. The histograms of the 5-tuples found in the sequences are shown in Fig. 3. With no noise, the sequence is a predictable alternation of inhibitory '0' and excitatory '1' states. By introducing a small amount of noise ($\epsilon = 0.02$), different 5-tuples begin to emerge as shown in the middle histogram in Fig. 3, indicating that the sequence is no longer completely predictable. However, it is only with a high level of noise ($\epsilon = 0.1$) that the sequence of states becomes random, approximating a binomial distribution: in the bottom histogram of Fig. 3 all 5-tuples are similarly represented.

These findings raise two important questions. First, why does the absence of noise lead to the alternation of inhibitory '0' and excitatory '1' states? Second, what is the mechanism by which noise disrupts the regular pattern? To provide answers, the weight dynamics during an anti-Hebbian phase are shown in Fig. 4. Fig. 4*A* shows the pure anti-Hebbian dynamics without noise. At each step, the stronger weight is decreased and the weaker weight is increased. This alternation proceeds regularly during the anti-Hebbian phase while the amplitude of the weight change decreases progressively. On the other hand, Fig. 4*B* shows that the regular alternation of weights is occasionally disrupted by noise, making it unpredictable which weight will emerge to be the stronger.

Fig. 4*A* shows that the high weight before the anti-Hebbian phase becomes low afterwards because the weights change an odd number of times. Were there an even number of updates, the initial high weight would return to being high after the anti-Hebbian phase. Therefore, applying random durations to the anti-Hebbian phases appears to be another way to introduce the necessary variation in the system to produce unpredictable reconfigurations. However, when a network is composed of many neurons undergoing the same modulatory signal, the variation in duration of such signals affects all neurons *simultaneously* and therefore produces an unwanted correlation. To highlight this point, the algorithm was run again on a network with one input neuron and *five output neurons*, each of which is connected as in Fig. 1. In this experiment, the durations of the anti-Hebbian phases are randomly set to an odd or even number of steps with

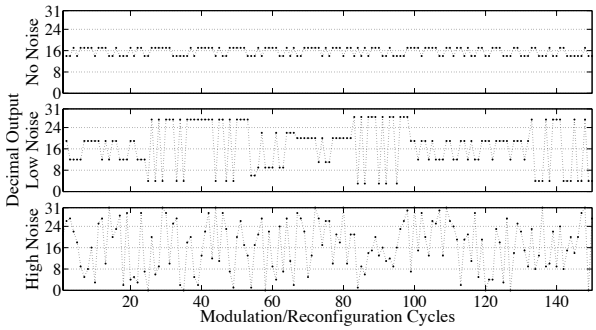

Figure 5: States of a five-neuron network undergoing 150 reconfiguration phases with variable duration of modulation. Each discrete value on the y-axes represents a different state mapped into the decimal interval 0-31. The graphs show that in the experiment with no noise (top graph), the network can reconfigure itself, but assumes only two complementary states, 01110 and 10001 (14 and 17 in decimal representation), demonstrating why changing the duration of modulation alone is not enough to encourage full exploration. With a low level of noise (middle graph), the network expresses a higher level of variation, assuming different configurations, but switches mostly between complementary states. With high noise (bottom graph), the network reconfigures itself to a new input-output mapping at each modulation cycle. Such randomness is key when anti-Hebbian plasticity is used to explore the space of network states.

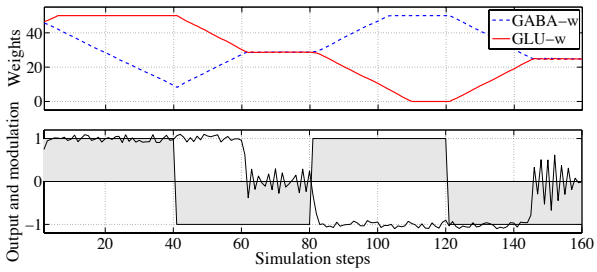

Figure 6: Weights and output dynamics with a large saturation value. Modulation is represented here as a shaded area overlapping the output signal. The weights now take longer than in the experiment of Fig. 2 to increase and decrease to and from the saturation value. This delay implies that weight reconfiguration can be reached only after a prolonged exposure to negative modulation. Thus this feature allows for a greater weight stability in the presence of stochastic noisy modulation, such as brief bursts of negative modulation over a trend of positive values.

probability 0.5. Because there are five output neurons, each modulation phase now results in not one but five output values. The change of those values over modulation phases describes the changing pattern of configuration states. Fig. 5 shows the results of the simulation with 150 modulation phases and three noise levels as before. The plots indicate that the variable length modulation makes the state change unpredictable, but the network does not visit all possible configuration states unless different noise dynamics affect each pathway independently.

In conclusion, the capability to reconfigure the network proves to be related to the level of noise. Only with a sufficient level of noise is the network capable of jumping to a completely random state at each reconfiguration cycle. Therefore, anti-Hebbian plasticity enforced by negative modulation expresses its full potential to reconfigure pathways, and consequently in the exploration of all network states, only when complemented with a sufficiently high level of noise.

3.3. The role of saturation

While noise is essential during anti-Hebbian reconfiguration phases, the maximum weight, or saturation value λ, is the stable state of weights after some duration of positive modulation. The maximum weight value must be large enough to overcome noise on neural transmission and small enough to make the updates significant in the time window of a simulation. The combined effect of the plasticity rate and the saturation value determines the time to convergence from mean value to saturation and vice versa.

In a new simulation, the saturation value λ was set to the large value of 50. Fig. 6 shows the slow climb and descent of weights after modulation changes. Such a long convergence time allows the network to preserve its state for a longer time in the face of negative modulation. It is important to note that a slower decay time of weights does not imply significantly longer-lasting memory, as is generally assumed elsewhere. Rather, the slow weight decay represents an index of greater inertia against change in the face of stochastic signals with hidden averages. A pathway converges to a neutral value (neither excitatory nor inhibitory) only when the modulation is negative on average over a long period. Thus the plasticity rule has the property of detecting average trends in a noisy signal. This property can be beneficial for addressing real-world noisy reward contingencies (Montague et al., 1995; Niv et al., 2002) and it is therefore further analyzed later in section 3.5.1.

3.4. From modulation to Braitenberg vehicles

An important implication of the previous simulations is that if the mode (i.e. excitatory or inhibitory) of a pathway can be associated with an *action*, then initially random actions are reinforced in the presence of positive modulation and extinguished in the presence of negative modulation. This section explains how this principle can be exploited in a simulated behavioral test. The chosen experiment is a navigation task in which an agent moves inside a closed arena and occasionally encounters different types of objects with different properties.

The geometrical properties of the environment and of the agent are similar to those of Braitenberg vehicles (Braitenberg, 1984). Braitenberg vehicles are moving robots capable of simple navigation and equipped with symmetrical sensors in the front. Thanks to the symmetrical configuration of sensors and the difference in speed between two parallel wheels, Braitenberg vehicles are capable of turning, avoiding or approaching objects, and displaying seemingly emotional behavior, whose emergence is

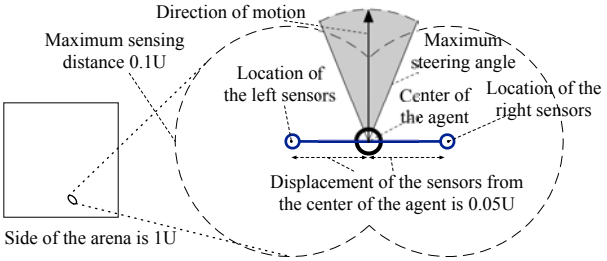

Figure 7: Navigating agent. The agent navigates in a square arena of unitary size (1U) with a constant speed of 0.01U/time step. The sensors are located symmetrically to the left and to the right of the agent. The activation of each omnidirectional sensor is inversely proportional to the distance of the object. The maximum sensor value is 1 when the distance between the sensor and an object is zero. Each sensor decreases its activation and reaches zero at the maximum sensing distance. The vehicle is also equipped with a unique modulation sensor (not shown). In this way, the amount of modulation is an input to the circuit regulated by the effect of behavior in the environment. The difference in rotation speed between the left and right wheels of a traditional Braitenberg vehicle is represented here by the activation of one output neuron whose value determines the steering angle in the range $[-18°, 18°]$.

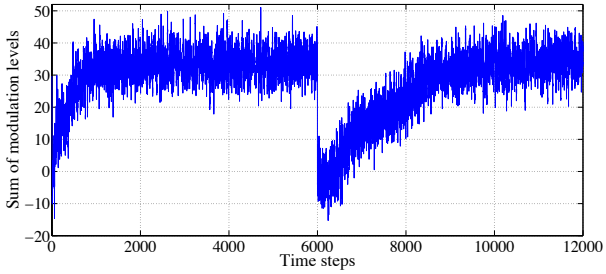

Figure 8: Modulation level through one lifetime, averaged over 200 lifetimes. Values, initially negative, tend to become positive over the early lifetime of the agent. When the modulation policy is changed at step 6,000, the agent suddenly receives negative modulation from the objects that earlier gave it positive modulation. However, the agent progressively modifies its behavior such that the sign of modulation returns to positive again. During the first half of the simulation, high modulation is reached more quickly than in the second half because initially the agent starts with uniform equal weights and therefore displays weak unbiased behavioral responses. In the second half of its lifetime the agent is first unlearning the behaviors that induce negative modulation before learning the new correct ones.

defined by Braitenberg as *synthetic psychology*. The vehicles are particularly suited to demonstrating the properties of the reconfigure-and-saturate rule because such vehicles describe relationships between neural wiring and higher-level behaviors. The simulations in this section focus in particular on Braitenberg vehicles 2a and 2b, described in the section "Fear and Aggression" (Braitenberg, 1984, Pages 6-9), which can either escape or attack an object according to the prewired input-output connections.

To reproduce the geometric placement of sensors in Braitenberg vehicles, in the present simulations each sensor is present in pairs located symmetrically to the left and to the right sides of the agent as shown in Fig. 7. The agent is equipped with wall, type-A object and type-B object proximity sensors, resulting in three pairs of sensors projecting their pathways (as in Fig. 1) onto one afferent output neuron. The activity of the output neuron sets the difference in speed between the left and right wheels of a typical Braitenberg vehicle, effectively determining the steering direction.

The agent has no a priori knowledge of the world, except for the capability of detecting signals. All signals, i.e. walls and objects, are processed equally and carry no initial meaning to the agent. However, the world-agent interaction causes the agent to receive different levels of modulation according to the circumstances. The purpose of devising modulation policies is to observe self-organizing behavior that derives from the reconfigure-and-saturate Hebbian plasticity under different environmental conditions. Instead of wiring the vehicle in a specific manner and observing the behavior that emerges from that wiring, as done by Braitenberg (1984), the proposed plasticity rule solves the reverse problem of allowing for the emergence of both wiring and behavior from a given modulation policy.

To signify that hitting walls is undesirable, approaching and impacting a wall causes negative modulation. Type-A objects also cause negative modulation on approach, but positive modulation while fleeing. Type-B objects are the opposite of type-A objects: approaching causes positive modulation while fleeing causes negative modulation. The numerical values of modulation are given in Appendix B. When the agent moves centrally over an object, the object is "eaten" or "destroyed". Eaten objects regenerate once the agent has moved outside the sensing radius. The modulation policy for walls does not change throughout the simulation. On the other hand, type-A and type-B objects can exchange their modulation policies from time to time. Relearning and rememorization of an acquired behavior is an important property of animal behavior (Staddon, 1983), also called *reversal* (Hasselmo et al., 2002). Such property is tested by changing the modulation policy during one simulation as described.

3.4.1. Simulation results

In a first experiment, the agent was tested over 200 *lifetimes* with a duration of 12,000 steps each. Each lifetime started with all weights reset to their middle value (5 in this setting). To ensure varied initial conditions, the agent started each lifetime at the position where it ended the previous one and the order of policies alternated across lifetimes. The modulation policies for type-A and type-B objects were exchanged halfway through each lifetime (i.e. at step 6,000). The modulation received by the agent averaged over 200 lifetimes is plotted in Fig. 8.

The plot indicates that the modulation level, which on average is negative at first, tends to increase during the lifetime. The change in modulation policy (at step 6,000) causes the modulation to drop to negative values, which then grow again to positive values. This trend indicates

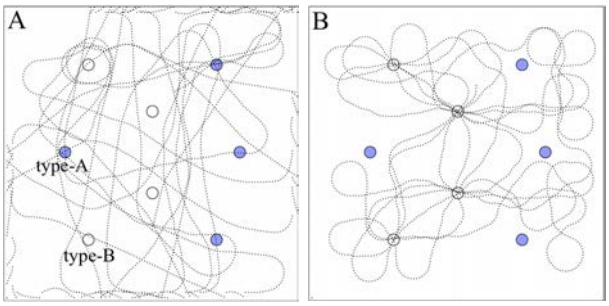

Figure 9: Samples of navigation paths. The path during a navigation of 2,000 time steps is shown in (A). For this particular plot, the weights are reset every 100 time steps to show the typical navigation pattern during the initial 100 time steps of a simulation. The agent hits walls and displays weak or uncertain responses to objects. In (B), 2,000 time steps are sampled *after* the learned behavior is stable and without weight reset. The agent displays fearful behaviors towards walls and type-A objects and aggressive behavior towards type-B objects; see also the supporting video at http://andrea.soltoggio.net/rec-sat.

that the behavior of the agent adapts continuously by extinguishing behaviors leading to negative modulation and reinforcing behaviors leading to positive modulation. Direct observation of the agent's behavior indeed confirms that the agent learns on its own to avoid walls, is attracted by one type of object and escapes the other type of object. Navigation samples are shown in Fig. 9.

The emergence of the agent's behavior during different stages of the simulation was analyzed in increasingly longer lifetimes of 100, 500 and 1,000 time steps. The location of the agent over many such lifetimes is plotted in the temperature graph in Fig. 10*A*-*C*. While at first the agent appears to navigate uniformly across the arena (Fig. 10*A*), as learning takes place, the agent becomes better at distinguishing significant features in the environment. Wall avoidance and frequent visits to one type of object are established at first (Fig. 10*B*-*C*). Fig. 10*D*-*E* show the agent's location after learning. The agent displays precise navigation patterns and preferences for one type of object. Those preferences are inverted when object types exchange modulation policies, as shown in Fig. 10*E*. The video of simulation provided as support material also shows that the agent's behaviors are equally acquired when objects slowly move across the area. The agent also relearns quickly a correct neural wiring when the motor output is suddenly inverted.

A key point in this study is that the acquisition of behaviors as displayed in Fig. 9*A*-*B* and 10*A*-*E* is a direct consequence of the weight dynamics induced by the reconfigure-and-saturate rule. To demonstrate this causality, the values of the six GLU weights from the inputs to the output are plotted in Fig. 11. The plot shows that weights diverge with time and reach high or low values. However, when the modulation policy changes, some weights undergo reconfigurations. Interestingly, two of the weights, namely those corresponding to the wall sensors, do not change. Such stability makes sense: while the modulation policy changes for type-A and type-B objects, it remains unchanged for walls. The simulation shows that the network can reconfigure those connections involved in the tasks that need to be relearned, but leaves unchanged those weights responsible for correct behaviors such as wall avoidance. This result means that the agent preserves memories related to particular stimuli even when it is coping with other orthogonal problems, as long as they are independent, such as those of facing type-A or type-B objects. The weights related to wall stimuli would reconfigure if the modulation policy governing walls changed.

It is interesting to note that the weight reconfiguration at the moment of policy change does not occur at the same speed for the weights from type-A and type-B sensors. In fact, the change is faster for type-B weights, i.e. those weights that previously determined an object-seeking behavior. When the modulation policy switches, the agent continues to target the previously positive objects (now negatively modulating), thus collecting negative modulation and causing the behavior to reverse at a faster rate. In contrast, type-A objects are avoided at the moment of policy switch with a consequent minimization of modulatory signal and longer learning time. This additional result is a simple demonstration that not only does learning affect behavior, but that behavior itself, by determining seeking or avoidance of stimuli, in turn affects learning. As an example, one final weight configuration during a static modulation policy is shown in Fig. 12.

The experiments presented in this section show that the "fearful" and "aggressive" behaviors described by Braitenberg (1984) can emerge in the simulated agent solely from the environmental modulation policies acting on the reconfigure-and-saturate Hebbian rule. In other words, the proposed Hebbian rule autonomously finds the connectivity and learns the behaviors of Braitenberg vehicles exclusively from the consequences of the agent's actions.

3.5. Action selection in multiple-input and multiple-output networks

Reconfigure-and-saturate Hebbian plasticity can be tested on networks with a higher number of inputs and outputs. Such networks can potentially solve an arbitrary number of problems, in which each individual input corresponds to one particular problem and the output vector represent a solution to the selected problem. Assume in particular that one problem is selected by activating one input, while one unique output vector, i.e. a pattern of activations across all outputs, represents the correct solution. Consistently with the previous experiments, the response from the environment is devised such that one particular output vector causes positive modulation while all the other output vectors cause negative modulation. The unique output vector that causes positive modulation represents the solution to the selected problem, while the other possible output vectors represent wrong answers. The purpose is to test that (1) the network can find the

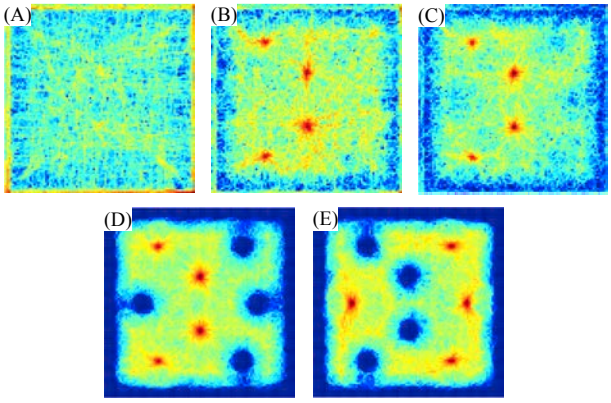

Figure 10: Average location of the agent in the area during different phases of learning. The high-temperature areas indicate a more frequent presence of the agent with respect to the low-temperature areas. (A) In the early simulation phase (100 steps repeated 1,000 times), the agent does not show distinctive navigation patterns and collides frequently with walls. (B) Over a longer lifetime (500 steps repeated 200 times) a wall-avoidance navigation pattern starts to emerge as well as frequent visits to one type of object. (C) Over an even longer lifetime (1,000 steps) good wall avoidance and visits to one type of object are established. (D) After 3,000 steps, the agent was left running for other 197,000 steps. The agent displays good wall avoidance, frequently visits type-B objects and avoids type-A objects. (E) To show the agent's behavior after the rewards for objects are switched, the agent's location is shown from step 203,000 to step 400,000. The agent reverses its preferences and now visits type-A object locations and correctly avoids type-B objects. A gray-scale version of the temperature graphs is included in Appendix B.

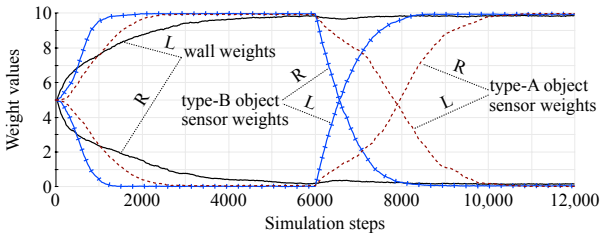

Figure 11: Average GLU weights during odd-numbered lives (encompassing 100 lifetimes) starting with policy 1 and ending with policy 2. The labels "L" and "R" indicate weights connected to the left and right side of the agent.

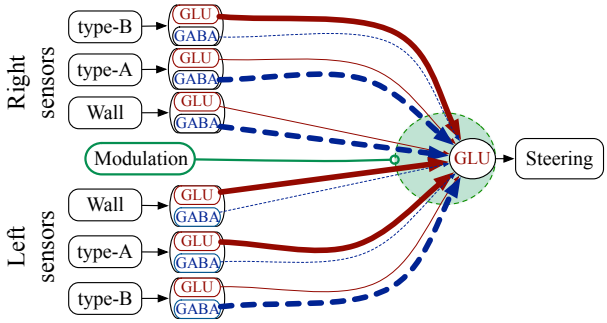

Figure 12: Twelve weights (six pathways) connecting the sensory inputs to the output with policy 1. The thickness of the line denotes the synaptic weight. The activation of the output neuron determines the turning strength such that inhibition results in left turns and excitation in right turns. This weight configuration yields a network in which stimuli on the left wall sensor excite the output, thereby causing the agent to steer to the right. Conversely, stimuli on the right wall sensor inhibit the output, causing the agent to steer to the left. The same strategy was learned when facing type-A stimuli: the agent steers away from them. Stimuli from type-B objects instead elicit an opposite response in the output, which leads the agent to point directly at those objects and hit them frontally.

unique weight configuration from one input to the outputs that causes positive modulation and (2) that such a network can have multiple inputs representing different problems and that the same output neurons can provide solutions to different problems. Fig. 13 illustrates the network.

The problem is structured as an n-armed bandit problem (Sutton and Barto, 1998), a test for reward learning algorithms in which n *arms*, also called *choices* or *options*, are associated with different rewards. In such problems, the amount of reward returned by each arm is initially unknown to the agent. The agent chooses one arm and immediately receives a reward that reflects the value of its choice. The task is to explore the arms and adopt a selection pattern that maximizes the total reward in the long term. The literature presents different types of bandit problem characterized by having static or dynamic reward policies, deterministic or stochastic rewards and, when dynamic, by the functions governing the changes in reward policy (Sutton and Barto, 1998).

The problems presented in this section have dynamic, stochastic rewards, which makes them more difficult than when they have static deterministic rewards. However, to focus on the plasticity rule, the mapping between rewards and modulation is preset by assigning negative modulation to all suboptimal arms (output vectors) and positive modulation to the only optimal arm. In other words, the estimation of the relative values of rewards, as traditionally done by algorithms solving n-armed bandit problems, and the identification of the hidden Markov processes underlying the reward policies are not the focus of this experiment. This fact highlights that the reconfigure-and-saturate rule does not prescribe particular

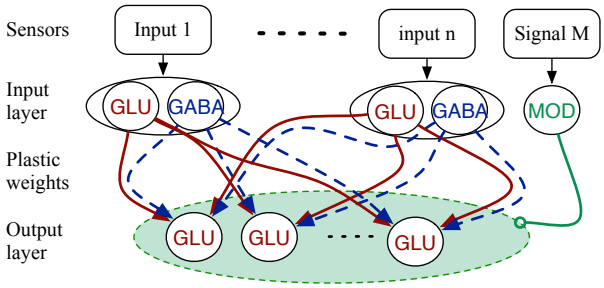

Figure 13: A multiple-input, multiple-output network. Each input represents one problem. The pattern of output values encodes the chosen answer.

exploratory/exploitative regimes. In particular, the rule explores the search space randomly and exploits continuously positively modulating action. Thus the plasticity rule is tested specifically for its capability to reinforce actions leading to positive modulation in this particular type of problem, and not as a general reinforcement learning algorithm.

The experiment is implemented by allowing h output neurons to encode $2^h = n$ binary patterns. A total of m inputs select one of m different problems. The test is performed by activating one of the m inputs at a time, observing the h output values and comparing them with a target binary sequence. If the output neurons display activation signs matching the target sequence, the problem is solved and the network receives a positive modulation of 1, i.e. $\eta(t) = 1$ during that time step. If any of the activation signs differ from the target sequence, a negative modulation is given proportionally to the number of errors up to -1 when no output matches the target.

The assumption that the inputs representing different problems must display phasic activity, i.e. be high, *one at a time*, ensures soundness. That is, in this domain, allowing more than one input to be active simultaneously would mean asking the output vector to answer two problems, possibly with different answers, at the same time. In analogy to the behavioral experiment with the Braitenberg vehicle, it would be similar to placing one object to flee and one to attack in the same place: in this condition, no correct behavioral response exists. However, it is important to note that while only one input may display phasic activity, the other inputs still display continuous tonic noisy activity, i.e. they continue to send background noisy signals to the outputs.

A first test includes three bandit problems of eight arms each, represented in binary by three output neurons. In a second test, six problems of 64 arms each are tested with a network of six inputs by six outputs. One lifetime is composed of four learning sessions: the target patterns of each bandit problem are randomly initialized and changed to new random patterns three times during a lifetime. The aim is to show that the network can learn the initial correct pattern for each of the problems, maintain such outputs while still valid and relearn new outputs when the target

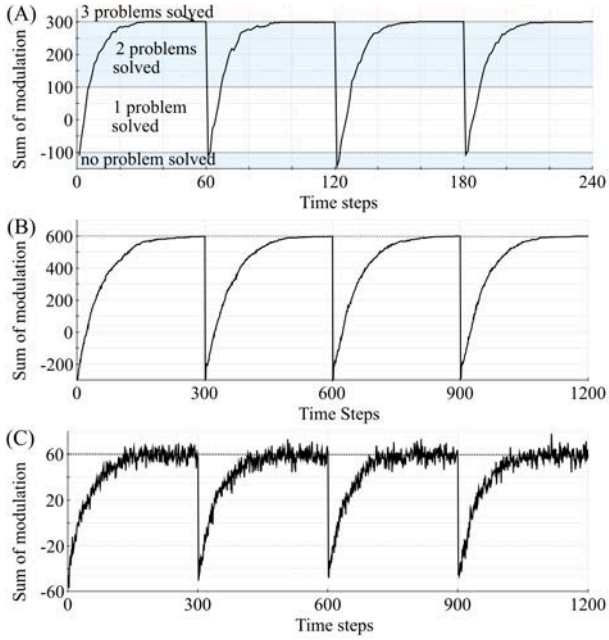

Figure 14: (A) Solving three problems with eight arms. One learning session, i.e. one phase during which the output target does not change, lasts 60 steps. One lifetime includes four learning sessions. When solved correctly, each problem produces a unitary modulation value. During one scan of all problems, the sum of modulation values thus can be at best 3. The modulation over 100 lifetimes is then summed. The value of 300 is reached when the solution is found for each of the three problems in all 100 simulations. When the maximum modulation level is reached, the network continues to provide correct answers to all problems as long as the target pattern is not changed. (B) Solving six problems with six outputs, i.e. 64 arms. To accommodate the large number of arms, the learning session is lengthened to 300 steps. The six problems were solved correctly in all of the 100 simulations. In other words, the network found consistently in 100 independent simulations the unique correct configuration of the 36 pathways in each of the four learning sessions. (C) Solving three problems with eight arms with highly stochastic modulation values. The sum of the modulation over the 100 runs approaches the value 60 on average (marked in the plot) before the target switch. The implication is that even with highly stochastic modulation, the network finds consistently the correct weight configurations to output the correct answer to the three problems.

solutions change. The simulation is performed by scanning all problems by activating all inputs one at a time for one simulation step. The simulation was repeated 100 times to assess robustness. The modulation received by the network while scanning each problem over all the 100 lifetimes was recorded to capture the dynamics of learning in this domain. The modulation values are shown in Fig. 14*A*. The modulation level increases consistently in all learning sessions, which means that the network changes the weights such that the output matches the target sequence. Once the target pattern is matched for all problems, modulation is maximized.

Fig. 14*B* shows the modulation values for the test with six problems and 64 arms. In this case, there are six unique target configurations, i.e. one for each problem, that are correctly matched by the output vector. This success indicates that even with a high number of arms, 63 of which

give negative modulation and only one of which gives positive modulation, the network modifies its outputs to find eventually a configuration that produces positive modulation.

This implementation does not consider a moving average of past rewards as in traditional solutions to n-armed bandit problems because the focus here is on the neural dynamics established by positive and negative modulation rather than on solving the reinforcement learning problem. Nevertheless, these dynamics can be interpreted as solving reward learning under the assumption that positive modulation represents higher-than-average reward. If such an assumption cannot be made, but there exists a stable association between outputs and modulation, the reconfigure-and-saturate rule will still drive the network towards outputs that causes positive modulation. This principle holds even if those dynamics do not constitute an optimal reinforcement learner. From this viewpoint, rather than *reinforcement learning* (Sutton and Barto, 1998), this algorithm mimics animal *operant reward learning* (Thorndike, 1911; Staddon, 1983), where actions leading to a successful outcome can be reinforced regardless of their optimality.

3.5.1. Dynamic stochastic modulation policies

Section 3.3, "The role of saturation", described the stable weight configuration as capable of detecting long-term hidden averages of stochastic modulatory processes. That claim is confirmed by performing an experiment with a highly stochastic modulation policy.

Stochasticity is implemented by assigning a modulation of 0.2 ± 0.5 to correct answers and -0.2 ± 0.5 to wrong answers. Thus, although correct answers provide on average positive modulation, they can occasionally give negative modulation and wrong answers can occasionally give positive modulation. Fig. 14C shows the modulation total in a three-input, three-output network over 100 runs. The solution of three problems over 100 runs results in a modulation of 60 on average. The plot indicates that this value is reached within 200 steps in all four learning sessions. To appreciate the difficulty of this problem, the modulation values and weight dynamics during one run are plotted in Fig. 15. The plot shows that the weights leave the saturation state frequently during the run. This phenomenon is due to the occasional negative modulation, as is evident from the first row of Fig. 15. However, the saturated weights have a sufficiently large value to overcome the temporary negative modulation without network reconfiguration. Thus the saturation value determines the *inertia* against change. A larger saturation value allows an even greater robustness to stochasticity of the modulatory signal; however, such robustness is counterbalanced by less readiness of adaptation when the hidden modulation average changes. In other words, as anticipated earlier in Section 3.3, the value of saturation appears to be a trade-off between readiness to change when the hidden reward average changes and robustness to stochastic rewards.

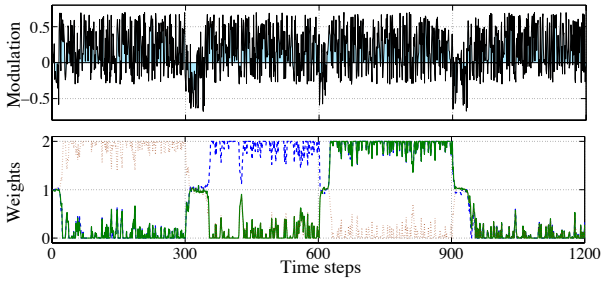

Figure 15: Modulation and weights with highly stochastic modulation. The reconfigure-and-saturate rule is capable of finding the weight configuration that maximize the modulation even when this signal is highly stochastic.

This experiment provides evidence that the learning dynamics extend to the stochastic case when the modulation is positive on average but can be occasionally negative. This behavioral robustness co-exists with the intrinsic adaptivity of the network and originates from the saturation dynamics of the model. The dynamic stochastic reward contingencies in this last experiment are often found in real scenarios (Montague et al., 1995; Niv et al., 2002). The capability of the reconfigure-and-saturate rule to extract the average of the modulation signal is therefore a fundamental property relevant to real-world applications.

4. Discussion and Future Work

The single-pathway experiment showed that one active input maps to either positive or negative stable activity in the output under positive modulation. Under negative modulation the output oscillates between positive and negative states. If modulation is interpreted as reward, this simple implementation of Hebbian plasticity augmented with modulation, neural noise and weight saturation represents a new most basic neural model of operant reward learning. The alternation of weight saturation and noisy weight oscillations, although deriving from simple correlation-decorrelation dynamics, can be interpreted as the equivalent behavioral manifestation of exploration and exploitation in reward learning. The analysis of both noise-driven dynamics and saturation values indicates that these are pivotal elements to implement the alternation between exhaustive exploration of the behavior space and stable exploitation. One novel aspect is the use of neural noise as the driving mechanism to reconfigure a negatively modulated network into a new set of weights with possibly different behavioral consequences. A second novel aspect is the use of weight saturation as a factor in determining the inertia against behavioral change, an important consideration when the reward is a stochastic noisy measure of a hidden average (Montague et al., 1995). The combination of these two elements in concert with modulated Hebbian plasticity represents the unique contribution of this study. The versatile dynamics in the first experiment

suggest that the model can be extended effectively to a simulated agent in a continuous simulated world.

The development of a Braitenberg vehicle further shows that the reconfigure-and-saturate rule can be successfully applied to a multi-input scenario where subproblems, e.g. type-A objects, type-B objects and walls, can be learned simultaneously and tackled separately and successfully by the simple one-output structure. Novel sensory information can be further plugged into the neural structure at any time. Noisy fluctuations of weights lead to initial behavioral responses that then cause a modulatory type of feedback. These behavioral responses, which are direct consequences of weights, are then reinforced or extinguished according to the modulation sign.

The multiple-input multiple-output experiment demonstrates that the plasticity rule presented in this paper can be applied to networks with many inputs and outputs without losing the learning property. The rule is thus applicable to learning multiple problems simultaneously, each of which may have a large search space. Interestingly, the modulation signal is global, and applies to all connections in the network; however, modulation affects the subpart of the network responsible for the current problem and leaves unaffected other weights because tonically active inputs, even when affected by noise, do not cause Hebbian updates. Learning with multiple outputs can also be employed to express one single value in a range, where the lower and upper bounds are expressed by all output neurons being inhibited or excited respectively, and intermediate values are expressed by the excitation of a subset of output neurons. Thus the limitation of saturating weights that cannot express intermediate values can be overcome by integrating one output signal over many output neurons. For example, the experiments in Section 3.5 showed that the network can learn one state out of 64, which can be interpreted as a 64-interval discretization of one continuous output value.

In both the navigation and bandit problems, specific stimuli are presented to the network for a short amount of time. In other words, after encountering a type-A object the navigating agent might proceed by meeting walls and type-B objects for a long time before encountering a type-A object again. Yet the reconfigure-and-saturate rule allows a quick synaptic update of the type-A sensory weights during the brief exposure to a type-A object. Those weights then remain unchanged while the agent deals with other problems (i.e. walls and type-B objects), thereby preserving the acquired memory for later use. Such a property of the reconfigure-and-saturate rule is essential in a multi-problem world in which different situations require different behaviors.

All experiments were structured such that the modulation is simultaneous in time to the actions that cause it. Therefore, the analogy of modulation with reward is possible under the assumption that the credit assignment problem is solved in the stimulus space (i.e. distinct stimulus types) but not with respect to time. In other words, the simple single-layer networks used in this paper cannot associate actions to modulation if the modulation is delivered later in time. Therefore, the one-layer structure does not implement temporal difference (TD) learning (Sutton and Barto, 1998); however, the credit assignment problem or sequence learning are not the focus of investigation in this paper. Other neural algorithms that use eligibility traces target those problems (Izhikevich, 2007; Soltoggio and Steil, 2012). The current study does not exclude the possibility that the reconfigure-and-saturate rule can be applied to multi-layer networks with longer time dynamics. One limitation in the current implementation of the rule is that exploration of the network states occurs randomly, while exploitation continues as long as positive modulation is perceived. Future work will focus on extending the rule to apply more sophisticated exploratory/exploitative regimes to solve a larger variety of reinforcement learning problems (Sutton and Barto, 1998). The one-layer structure is also capable of learning only linear problems (i.e. when the inputs can be treated as independent). However, it is important to appreciate that these limitations are inherent only in the one-layer structures analyzed in this paper. Just as the perceptron learning rule ultimately led to backpropagation (Werbos, 1974; Russell and Norvig, 2003), the potential of a new fundamental learning property is often first established in single-layer networks and later generalized.

An interesting research direction in this context is the combination of the proposed learning structure with reservoir recurrent networks (Maass and Markram, 2004) to complement nonlinear computation with modulation-driven learning. In particular, a large pool of randomly connected neurons, i.e. the reservoir, can potentially implement the nonlinear preprocessing stage from which the linear multiple-input multiple-output learning structure of Fig. 13 can read signals. In such a case, provided that the reservoir has rich enough dynamics, the plasticity rule on the linear output can also learn solutions to nonlinear problems. The reconfigure-and-saturate Hebbian rule can potentially also be integrated in larger artificial networks designed by neuroevolution (Stanley and Miikkulainen, 2002; Stanley et al., 2009). Simulated evolution can explore the application of the rule either at a global or at a local scale in relation to various learning scenarios.

One other exciting prospect is to apply the plasticity rule within large recurrent networks to reinforce or extinguish oscillatory attractors. Oscillation in neural circuitry is believed to play a crucial role both in pattern generation (Marder, 1996; Dickinson, 2006) and higher cognition (Buzaki, 2006). Future studies can investigate the dynamic behavior of completely or locally modulated recurrent networks as substrates for learning behavioral responses.

Finally, it is important to note that the saturate-and-configure Hebbian rule is not intended to predict biological neural dynamics. Nevertheless, the use of noise as a driving exploratory mechanism and saturation as a stable

state for exploitation hints at the potential for new interpretations and hypotheses for the corresponding biological dynamics. Indeed, recent evidence suggests that neural noise in the central nervous system is responsible for trial-to-trial variability (Faisal et al., 2008). A research question that emerges from the current study is whether the noise-induced variability in biological systems can be also responsible for behavioral exploration. Likewise, the stable exploitation of acquired neural functions, such as the consistent application of a skill, might be expressed through high-weight pathways that encode clear mappings among neurons. Just as the strength of the saturated synapse in the this study represents a level of resistance to the reversal of previously acquired behaviors (Hasselmo et al., 2002), a similar relationship might be also present in biological networks between the strength of pathways and the stability of behaviors.

5. Conclusion

A central insight of this article is that the traditionally undesirable weight growth of the Hebbian rule can in fact *benefit* a model of synaptic plasticity. In effect, such auto-correlative dynamics can amplify a random initial behavior when it proves beneficial to obtaining a reward. In the simple model of this paper, when negative modulation is registered, autocorrelation is reversed into decorrelation or anti-Hebbian plasticity. Such environment-driven alternation of Hebbian and anti-Hebbian plasticity, when complemented with sufficient neural noise and saturation boundaries, establishes a plasticity model that implements *exploratory* and *exploitative* behaviors. This model in turn results in simulated *operant reward learning* at the system level. The rule, effectively the simplest form of modulated Hebbian plasticity augmented with noise and saturation, produces exploration and exploitation with a simpler and more essential mechanism than those in previous studies of reward-modulated Hebbian plasticity.

The model was tested on a navigation problem in which not only wall avoidance was achieved, but also the "fearful" and "aggressive" behaviors of Braitenberg vehicles emerged purely from environmental reward policies. Exploratory and exploitative behaviors were thus achieved with the most basic Hebbian model. While such learning problems as those analyzed in this paper have been employed for decades to test the bottom-up emergence of intelligent behaviors in fields such as artificial life (Langton, 1990) and evolutionary robotics (Floreano and Nolfi, 2004; Floreano and Mattiussi, 2008), the results presented here show that very small structures can serve as exemplary paradigms of adaptive behavior (Staddon, 1983). The final experiment suggests that this plasticity rule maintains its problem-solving properties in multiple-input multiple output-networks in the face of highly stochastic signals, an important capability for the real world. All experiments showed that brief but significant stimuli are processed to capture and retain relevant information indefinitely. Such a property particularly fits real-world scenarios in which fast memorization of stimuli is often required for later reuse.

In summary, this study indicates that Hebbian plasticity, when augmented with neuromodulation, neural noise and weight saturation, acquires key dynamics that link plasticity to adaptive behavior and facilitates learning from short-lived but relevant events. With this model, the dynamics of Hebbian synaptic plasticity can be directly related to the learning, manifestation and memorization of behavior.

Acknowledgment

Special thanks are given to Christian Emmerich, Ben H. Jones, Andre Lemme, Klaus Neumann, Arne Nordmann, Natalja Prokoptsova, Felix Reinhart, Sebastian Risi, Matthias Rolf, Maha Salem and Jochen Steil for valuable comments on earlier drafts of this paper. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013 Challenge 2 Cognitive Systems, Interaction, Robotics under grant agreement No 248311 - AMARSi.

Appendix A. Spiking neuron simulation

The reconfigure-and-saturate Hebbian dynamics hold equally with a simple spiking neuron model (Wilson, 1999; Maass and Bishop, 1999; Gerstner and Kistler, 2002b). The membrane potential u can be computed as a leaky integrate-and-fire model Gerstner and Kistler (2002b), which can be expressed in discrete time as

$$u_i(t+1) = u_i(t) + \frac{1}{\tau}\Big(-u_i(t) + \sum_{j=1}^{i=n} w_{ji}(t)\cdot v_j(t)\Big) \quad \text{(A.1)}$$

where τ is the time constant that determines the rate of the leak, here set to the value 30. In the current simulation, the firing threshold is set to zero. For positive values of u, a spike is emitted with probability 0.5. Following a spike, u is reset to zero and the neuron has a resting time of one step during which it cannot spike. To detect positive and negative correlations in the plasticity rule, the relevant constraint here is to assume a positive output value for a spiking time step and a negative output value for non-spiking time steps, which are set respectively to 1 and -0.1. The plasticity rule remains unchanged.

The one-input one-output pathway experiment was run in an additional experiment with the spiking neural dynamics. Fig. A.16 shows that the weight change is similar to that of the rate-based model. The mapping from input to output neuron is expressed by the frequency of spikes. When the input inhibits the output, the output neuron does not spike. When a prevailing GLU weight is established, the output neuron spikes with high frequency. Interestingly, the negatively modulated phases are characterized by a residual level of spiking, which is neither

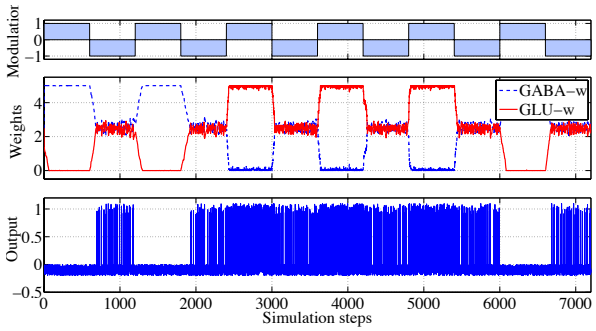

Figure A.16: Simulation of switching modulation with spiking neurons. The first row shows the imposed modulation value. The second row shows the values of the excitatory and inhibitory weights. Finally, the third row shows the spiking pattern of the output neuron. The absence of spikes indicates that the input causes inhibition of the output. A high spiking frequency, e.g. between steps 2,500 and 3,000, indicates that the input causes excitation of the output. When the modulation is negative, occasional spikes characterize the output.

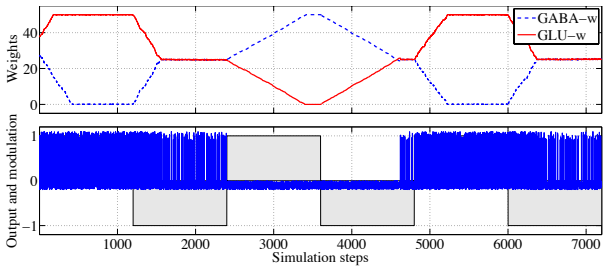

Figure A.17: Simulation of switching modulation with spiking neurons and a high saturation value $\lambda = 50$. Similarly to the simulation with the rate-based model in Fig. 5 in the main text, the network reaches the reconfigure dynamics only after a certain duration of negative modulation.

full activity nor silence and can be interpreted as the tonic firing rate. This frequency can be seen as an uncertain or neutral state of activation.

Fig. A.17 shows the weight dynamics and output pattern with a high saturation value $\lambda = 50$, which are similar to Fig. 6 in the main text, and spiking neural dynamics.

Appendix B. Agent's simulation

Details of the navigating agent are reported in this section.

Appendix B.1. Agent, arena and policies

The square arena has a side of length one unit (1U). The bounding walls and the type-A and type-B objects are illustrated in Fig. B.18*A*. The agent navigates with a constant speed of 0.01U per simulation step. One neuron is used as output. However, the alternative use of two output neurons, determining the speed of two wheels, would allow for the synthesis of more behaviors, similar to those described in the section "Love" (Braitenberg, 1984, Pages 10-14). These additional behaviors become possible because the agent could also increase or decrease the speed

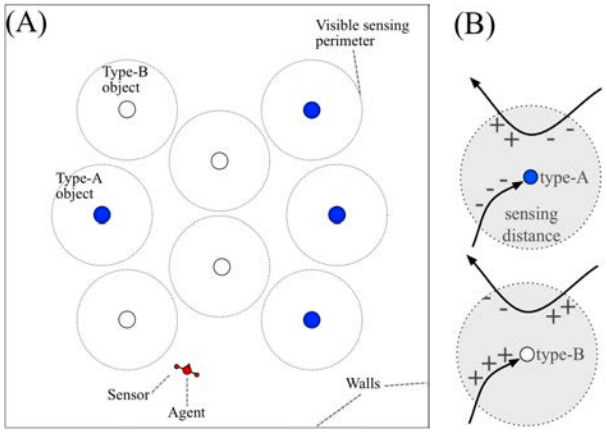

Figure B.18: (A) The arena, the bounding walls, and the type-A and type-B objects. (B) Modulation policies: plus signs indicate positive modulation and minus signs indicate negative modulation. When the agent meets an object by navigating over it, the object is removed temporarily, and reappears in the same location after the agent has left the sensing area.

$\forall d < \rho$	Policy 1	Policy 2
Walls	$-\Delta d/(2\kappa)$	$-\Delta d/(2\kappa)$
Type-A	$-\Delta d/\kappa$	$\Delta d/\kappa$
Type-B	$\Delta d/\kappa$	$-\Delta d/\kappa$

Table B.1: Values of η for the modulation policies in the agent's simulation. The maximum sensing distance κ (sensor range) is set to 0.1U in all experiments. The distance d of the agent from an object is used to compute $\Delta d = d(t) - d(t-1)$. The speed of the agent is (0.01U/step). The normalization implemented by dividing by κ means that the maximum modulation is 0.5 and 1.0 in absolute value for walls and objects, respectively.

while turning. Such an extension is a good candidate for increasing the set of possible behaviors in future work. It is also important to note that the simulation time is measured in time steps rather than in seconds; this convention implies that the speed of change of synaptic weights can be interpreted over different time scales.

When hitting a wall, the agent bounces back 0.02U and rotates 30° away from the wall. An object disappears when the agent navigates at a distance equal or less than 0.025U from it; the object reappears when the agent exits the sensing radius ρ of 0.1U. A graphical representation of modulation policies is illustrated in Fig. B.18*B*. Exact modulation values are reported in table B.1. Fig. B.19 is a gray-scale version of Fig. 10.

Appendix C. Computer simulations

The experiments presented in this paper were implemented and simulated in Matlab®. The complete Matlab code is provided as support material. All figures showing data and the data generated by the simulations can be reproduced exactly (using the same pseudo-random sequence) or qualitatively (using different pseudo-random sequences) with the provided Matlab code. The code is

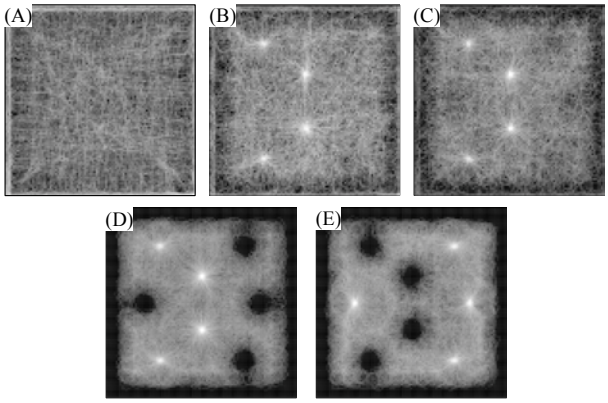

Figure B.19: Gray-scale version of Fig. 10 from the main text. The lighter areas indicate a more frequent presence of the agent with respect to the darker areas.

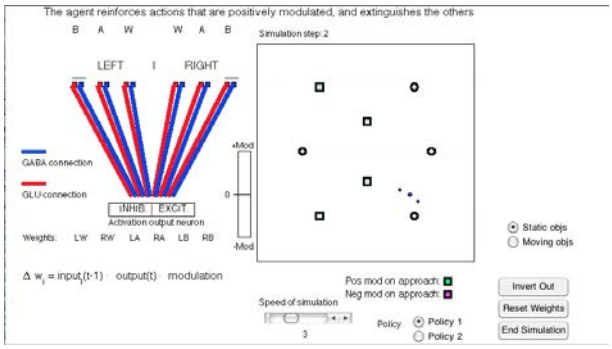

Figure C.20: Video showing the simulation process of the agent in the arena. The video and the source code to reproduce these simulations can be downloaded at http://andrea.soltoggio.net/rec-sat.

cross-platform, and does not require any particular installation procedure.

An example of one simulation of the agent, explaining the user interface and the phases for learning, is captured in a video provided with the study (Fig. C.20). The video and source code can be downloaded at this article's associate website `http://andrea.soltoggio.net/rec-sat`.

References

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

Alger, B. E., Nicoll, R. A., 1982. Feed-forward dendritic inhibition in rat hippocampal pyramidal cells studied in vitro. Journal of Physiology 328, 105–123.

Arkin, R. C., 1998. Behavior-Based Robotics. MIT Press.

Aston-Jones, G., Cohen, J. D., 2005. Adaptive Gain and the Role of the Locus Coeruleus-Norepinephine System in Optimal Performance. The Journal of Comparative Neurology 493, 99–110.

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

Baxter, D. A., Canavier, C. C., Clark, J. W., Byrne, J. H., 1999. Computational Model of the Serotonergic Modulation of Sensory Neurons in Aplysia. Journal of Neurophysiology 82, 1914–2935.

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

Berridge, K. C., Robinson, T. E., 1998. What is the role of dopamine in reward: hedonic impact, reward learing, or incentive salience? Brain Research Reviews 28, 309–369.

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

Bienenstock, L. E., Cooper, L. N., Munro, P. W., January 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. The Journal of Neuroscience 2 (1), 32–48.

Braitenberg, V., 1984. Vehicles: Experiments in Synthetic Psychology. The MIT Press.

Brembs, B., Lorenzetti, F. D., Reyes, F. D., Baxter, D. A., Byrne, J. H., 2002. Operant Reward Learning in Aplysia: Neuronal Correlates and Mechanisms. Science 296 (5573), 1706–1709.

Brown, T. H., Chapman, P. F., Kairiss, E. W., Keenan, C. L., 1988. Long-Term Synaptic Potentiation. Science 242, 724–728.

Brown, T. H., Kairiss, E. W., Keenan, C. L., 1990. Hebbian Synapse: Biophysical Mechanisms and Algorithms. Annual Review of Neuroscience 13, 475–511.

Buzaki, G., 2006. Rhythms of the brain. Oxford University Press.

Carew, T. J., Walters, E. T., Kandel, E. R., December 1981. Classical conditioning in a simple withdrawal reflex in Aplysia californica. The Journal of Neuroscience 1 (12), 1426–1437.

Clark, G. A., 2001. International Encyclopedia of the Social & Behavioural Sciences : Volume 26. Elsevier, Ch. Synaptic Efficacy, Regulation of, pp. 15371–15378.

Clark, G. A., Kandel, E. R., 1984. Branch-specific heterosynaptic facilitation in Aplysia siphon sensory cells. PNAS 81 (8), 2577–2581.

Cohen, M. X., 2008. Neurocomputational mechanisms of reinforcement-guided learning in humans: A review. Cognitive, Affective and Behavioral Neuroscience 8 (2), 113–125.

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

Cox, R. B., Krichmar, J. L., 2009. Neuromodulation as a robot controller: A brain inspired strategy for controlling autonomous robots. IEEE Robotics & Automation Magazine.

Dayan, P., Abbott, L. F., 2001. Theoretical Neuroscience. MIT Press Cambridge, MA, USA.

Dickinson, P. S., 2006. Neuromodulation of central pattern generators in invertebrates and vertebrates. Current Opinion in Neurobiology 16, 604–614.

Doya, K., 2002. Metalearning and neuromodulation. Neural Networks 15 (4-6), 495–506.

Doya, K., Uchibe, E., 2005. The Cyber Rodent Project: Exploration and Adaptive Mechanisms for Self-Preservation and Self-Reproduction. Adaptive Behavior 13 (2), 149–160.

Faisal, A. A., Selen, L. P. J., Wolpert, D. M., 2008. Noise in the nervous system. Nature Reviews Neuroscience.

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

Fellous, J.-M., Linster, C., 1998. Computational Models of Neuromodulation. Neural Computation 10, 771–805.

Finch, D. M., Tan, A. M., Isokawa-Akesson, M., 1988. Feedforward inhibition of the rat entorhinal cortex and subicular complex. The Journal of Neuroscience 8 (7), 2213–2226.

Floreano, D., Mattiussi, C., 2008. Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies (Intelligent Robotics and Autonomous Agents series). Intelligent Robotics and Autonomous Agents series. The MIT Press.

Floreano, D., Nolfi, S., 2004. Evolutionary Robotics. The MIT Press.

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

Frémaux, N., Sprekeler, H., Gerstner, W., 2010. Functional re-

quirements for reward-modulated spike-timing-dependent plasticity. The Journal of Neuroscience 30 (40), 13326–13337.
Fusi, S., Annunziato, M., Badoni, D., Salamon, A., Amit, D. J., 2000. Spike-Driven Synaptic Plasticity: Theory, Simulation, VLSI Implementation. Neural Computation 12 (10), 2227–2258.
Gerstner, W., Kistler, M. W., 2002a. Mathematical formulations of Hebbian learning. Biological Cybernetics 87, 404–415.
Gerstner, W., Kistler, M. W., August 2002b. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, Cambridge, UK.
Gil, M., DeMarco, R. J., Menzel, R., 2007. Learning reward expectations in honeybees. Learning and Memory 14, 291–496.
Giocomo, L. M., Hasselmo, M. E., 2007. Neuromodulation by Glutamate and Acetylcholine can Change Circuit Dynamics by Regulating the Relative Influence of Afferent Input and Excitatory Feedback. Molecular Neurobiology 36, 184–200.
Grossberg, S., 1976. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neuronal feature detectors. Biological Cybernetics 23, 121–134.
Gustafsson, B., Wigstroem, H., Abraham, W. C., Huang, Y.-Y., 1987. Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to a single volley synaptic potentials. The Journal of Neuroscience 7 (3), 774–780.
Hammer, M., November 1993. An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. Nature 366, 59–63.
Harris-Warrick, R. M., Marder, E., 1991. Modulation of neural networks for behavior. Annual Review of Neuroscience 14, 39–57.
Hasselmo, M. E., 1994. Runaway synaptic modification in models of cortex: Implications for Alzheimer's disease. Neural Networks 7 (1), 13–40.
Hasselmo, M. E., 1995. Neuromodulation and cortical function: modeling the physiological basis of behavior. Behavioural Brain Research 67, 1–27.
Hasselmo, M. E., 2006. The role of acetylcholine in learning and memory. Current Opinion in Neurobiology 16, 710–715.
Hasselmo, M. E., Bodelon, M. E., Wyble, B. P., 2002. A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. Neural Computation 14 (4), 793–817.
Hasselmo, M. E., Schnell, E., 1994. Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. Journal of Neuroscience 14 (6), 3898–3914.
Hebb, D. O., 1949. The Organization of Behavior: A Neuropsychological Theory. Wiley, New York.
Izhikevich, E. M., 2007. Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling. Cerebral Cortex 17, 2443–2452.
Kandel, E. R., Tauc, L., 1965. Heterosynaptic facilitation in neurones of the abdominal ganglion of Aplysia depilans. The Journal of Physiology 181, 1–27.
Kelso, S. R., Ganong, A. H., Brown, T. H., 1986. Hebbian Synapses in Hippocampus. PNAS 83 (14), 5326–5330.
Krichmar, J. L., 2008. The Neuromodulatory System: A Framework for Survival and Adaptive Behavior in a Challenging World. Adaptive Behavior 16, 385–399.
Langton, C., 1990. Computation at the edge of chaos: Phase-transitions and emergent computation. Ph.D. thesis, University of Michigan.
Legenstein, R., Chase, S. M., 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.
Levy, W. B., Steward, O., 1979. Synapses as associative memory elements in the hippocampal formation. Brain Research 175 (2), 233–245.
Levy, W. B., Steward, O., 1983. Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. Neuroscience 8 (4), 791–797.
Lisman, J., 1989. A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. PNAS 86, 9574–9578.
Ludvig, E. A., Sutton, R. S., Kehoe, E. J., 2008. Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. Neural Computation 20, 3034–3054.
Maass, W., Bishop, C. M., 1999. Pulsed Neural Networks. London; Cambridge, Mass. : MIT Press.
Maass, W., Markram, H., 2004. On the Computational Power of Recurrent Circuits of Spiking Neurons. Journal of Computer and System Sciences 69 (4), 593–616.
Marder, E., 1996. Neural modulation: Following your own rhythm. Current Biology 6 (2), 119–121.
Markram, H., Lübke, J., Frotscher, M., Sakmann, B., January 1997. Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. Science 275, 213–215.
Marr, D., 1969. A theory of cerebellar cortex. Journal of Physiology 202, 437–470.
McNaughton, B. L., Barnes, C. A., Rao, G., Rasmussen, M., 1986. Long-term enhancement of hippocampal synaptic transmission and the acquisition of spatial information. The Journal of Neuroscience 6 (2), 563–571.
Miller, K. D., Mackay, D. J. C., 1994. The Role of Constraints in Hebbian Learning. Neural Computation 6, 100–126.
Moldakarimov, S. B., Sejnowski, T. J., 2008. Concise Learing and Memory: The editor's selection. Elsevier, Ch. Neural Computation Theories of Learning.
Montague, P. R., Dayan, P., Person, C., Sejnowski, T. J., October 1995. Bee foraging in uncertain environments using predictive Hebbian learning. Nature 377, 725–728.
Montague, P. R., Dayan, P., Sejnowski, T. J., March 1996. A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. The Journal of Neuroscience 16 (5), 1936–1947.
Montague, P. R., Hyman, S. E., Cohen, J. D., 2004. Computational roles for dopamine in behavioural control. Nature 4, 2–9.
Niv, Y., Joel, D., Meilijson, I., Ruppin, E., 2002. Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviours. Adaptive Behavior 10 (1), 5–24.
Oja, E., November 1982. Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology 15 (3), 267–273.
Pennartz, C. M. A., 1996. The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. Brain Research Reviews 21, 219–245.
Pennartz, C. M. A., 1997. Reinforcement Learning by Hebbian Synapses with Adaptive Threshold. Neuroscience 81 (2), 303–319.
Pennartz, C. M. A., Kitai, S. T., September 1991. Hippocampal inputs to identified neurons in an in vitro slice preparation of the rat nucleus accumbens: evidence for feed-forward inhibition. The Journal of Neuroscience 11 (9), 2838–2847.
Pfeiffer, M., Nessler, B., Douglas, R. J., Maass, W., 2010. Reward-modulated Hebbian Learning of Decision Making. Neural Computation 22, 1–46.
Porr, B., Wörgötter, F., 2006. Strongly Improved Stability and Faster Convergence of Temporal Sequence Learning by Using Input Correlation Only. Neural Computation 18, 1380–1412.
Porr, B., Wörgötter, F., 2007. Learning with Relevance: Using a third factor to stabilize Hebbian learning. Neural Computation 19 (10), 2694–2719.
Rauschecker, J. P., Singer, W., 1981. The effects of early visual experience on the cat's visual cortex and their possible explanation by Hebb synapses. Journal of Physiology 310, 215–239.
Redgrave, P., Gurney, K., Reynolds, J., 2008. What is reinforced by phasic dopamine signals? Brain Research Reviews 58, 322–339.
Russell, S., Norvig, P., 2003. Artificial Intelligence: A Modern Approach, 2nd Edition. Prentice Hall.
Schultz, W., 2006. Behavioural Theories and the Neurophysiology of Reward. Annual Review of Psychology 57, 87–115.
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, P. R., 1997. A Neural Substrate for Prediction and Reward. Science 275, 1593–1598.

Smith, T., Husbands, P., Philippides, A., O'Shea, M., 2002. Neuronal Plasticity and Temporal Adaptivity: GasNet Robot Control Networks. Adaptive Behavior 10, 161–183.

Soltoggio, A., Bullinaria, J. A., 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., Dürr, P., Mattiussi, C., Floreano, D., 2007. Evolving Neuromodulatory Topologies for Reinforcement Learning-like Problems. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2007.

Soltoggio, A., Steil, J. J., 2012. Solving the Distal Reward Problem with Rare Correlations. Neural Computation (under review).

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

Sporns, O., Alexander, W. H., 2002. Neuromodulation and plasticity in an autonomous robot. Neural Networks 15, 761–774.

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

Stanley, K. O., D'Ambrosio, D. B., Gauci, J. J., 2009. A Hypercube-Based Encoding for Evolving Large-Scale Neural Networks. Artificial Life.

Stanley, K. O., Miikkulainen, R., May 2002. Evolving Neural Networks through Augmenting Topologies. Evolutionary Computation 10 (2), 99–127.

Stent, G. S., 1973. A Physiological Mechanism for Hebb's Postulate of Learning. PNAS 70 (4), 997–1001.

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

Thorndike, E. L., 1911. Animal Intelligence. Macmillan.

Turrigiano, G. G., 2008. The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. Cell 135, 422–435.

van Rossum, M. C. W., Bi, G. Q., Turrigiano, G. G., 2000. Stable Hebbian Learning from Spike Timing-Dependent Plasticity. The Journal of Neuroscience 20 (23), 8812–8821.

Walters, E. T., Byrne, J. H., 1983. Associative Conditioning of Single Sensory Neurons Suggests a Cellular Mechanism for Learning. Science 219, 405–408.

Werbos, P. J., 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. thesis, Harvard University.

Wilson, H. R., 1999. Spikes, decisions, and actions : the dynamical foundations of neuroscience. Oxford : Oxford University Press.

Wise, R. A., Rompre, P. P., 1989. Brain dopamine and reward. Annual Review of Psychology 40, 191–225.