Neural Networks 151 (2022) 16–33

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Review

Exploration in neo-Hebbian reinforcement learning: Computational approaches to the exploration–exploitation balance with bio-inspired neural networks

Anthony Triche[*], Anthony S. Maida, Ashok Kumar

The Center for Advanced Computer Studies, University of Louisiana at Lafayette, 301 East Lewis Street, P.O. Box 43694, Lafayette, LA 70504-3694, United States of America

ARTICLE INFO

Article history:
Received 29 June 2021
Received in revised form 8 March 2022
Accepted 14 March 2022
Available online 23 March 2022

Keywords:
Reinforcement learning
Hebbian plasticity
STDP
Spiking neural networks
Exploration–exploitation balance

ABSTRACT

Recent theoretical and experimental works have connected Hebbian plasticity with the reinforcement learning (RL) paradigm, producing a class of trial-and-error learning in artificial neural networks known as neo-Hebbian plasticity. Inspired by the role of the neuromodulator dopamine in synaptic modification, neo-Hebbian RL methods extend unsupervised Hebbian learning rules with value-based modulation to selectively reinforce associations. This reinforcement allows for learning exploitative behaviors and produces RL models with strong biological plausibility. The review begins with coverage of fundamental concepts in rate- and spike-coded models. We introduce Hebbian correlation detection as a basis for modification of synaptic weighting and progress to neo-Hebbian RL models guided solely by extrinsic rewards. We then analyze state-of-the-art neo-Hebbian approaches to the exploration–exploitation balance under the RL paradigm, emphasizing works that employ additional mechanics to modulate that dynamic. Our review of neo-Hebbian RL methods in this context indicates substantial potential for novel improvements in exploratory learning, primarily through stronger incorporation of intrinsic motivators. We provide a number of research suggestions for this pursuit by drawing from modern theories and results in neuroscience and psychology. The exploration–exploitation balance is a central issue in RL research, and this review is the first to focus on it under the neo-Hebbian RL framework.

© 2022 Elsevier Ltd. All rights reserved.

Contents

[*] Corresponding author.
E-mail addresses: anthony.triche1@louisiana.edu (A. Triche), anthony.maida@louisiana.edu (A.S. Maida), ashok@louisiana.edu (A. Kumar).

https://doi.org/10.1016/j.neunet.2022.03.021
0893-6080/© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Reinforcement is a potent and biologically plausible paradigm for learning (Sutton & Barto, 2017), with promising results for artificial neural networks (ANNs) under a set of methods known as neo-Hebbian plasticity (Gerstner, Lehmann, Liakoni, Corneil, & Brea, 2018). For spiking neural networks (SNNs), neo-Hebbian learning is also commonly known as reward-modulated spike-timing-dependent plasticity (R-STDP) (Frémaux & Gerstner, 2016).

Hebbian learning comprises a class of two-factor formulations for unsupervised learning of synaptic weights based on coincident activity between pre- and post-synaptic units. For SNNs, unsupervised STDP implements Hebbian plasticity on the basis of near-coincident activity, allowing changes to synaptic strength to scale or reverse sign according to the temporal proximity and ordering of pre- and post-synaptic spike events.

Inspired by neuroscientific research on the neuromodulator dopamine, neo-Hebbian methods extend Hebbian plasticity through the incorporation of environmental rewards as a third factor to gate and modulate the scale of weight updates. Neo-Hebbian plasticity has been shown to enable semi-supervised (Gardner & Grüning, 2013; Hoerzer, Legenstein, & Maass, 2012; Mozafari, Ganjtabesh, Nowzari-Dalini, Thorpe, & Masquelier, 2018; Pogodin, Corneil, Seeholzer, Heng, & Gerstner, 2019) as well as temporal-difference (TD) learning in both spiking (Izhikevich, 2007) and non-spiking ANNs (Gordon, Dorfman, & Ahissar, 2013).

An increasing body of reinforcement learning (RL) research on the topic of neo-Hebbian learning has prompted a number of review articles (Feldman, 2012; Frémaux & Gerstner, 2016; Gerstner et al., 2018; Kuśmierz, Isomura, & Toyoizumi, 2017; Roelfsema & Holtmaat, 2018; Shouval, Wang, & Wittenberg, 2010; Tetzlaff, Kolodziejski, Markelic, & Wörgötter, 2012). These reviews range in generality and can be categorized to include works which: (i) cover neo-Hebbian learning across all of the major machine learning paradigms, (ii) examine the role of neo-Hebbian RL under broader contexts such as the consolidation of associative memories, and (iii) assess the relationship between specific neo-Hebbian formulations and their counterparts in computational RL theory.

There has been no substantive treatment on the exploration–exploitation balance in this context despite a growing interest in neo-Hebbian RL. This gap in the review literature is particularly opportune given an increasing volume of computational works expanding upon neo-Hebbian plasticity to integrate these concepts. The development of novel neo-Hebbian exploration methods may be further aided by drawing from a vast body of research in neuroscience and psychology on the causes and mechanics of exploratory behaviors in nature.

This review targets two primary objectives: (i) to elucidate the limitations of current neo-Hebbian RL methods which modulate exploration–exploitation dynamics through extended learning formulations; and (ii) to provide novel recommendations, derived from recent research in biological learning, which will enable future work in this domain to improve upon existing methods substantially. Proper coverage of the state–action space of an environment is essential not only to learning accurate estimates of the long-term value of context-specific behaviors but also to the ability of a learning agent to generalize between similar environments.

The construction of this review proceeds as follows: the remainder of Section 1 introduces topics pertinent to the scope of the review, presents motivations for pursuing their study, and discusses relevant contributions made by related reviews on neo-Hebbian learning; Section 2 provides an approachable introductory treatment on foundational neo-Hebbian methods, including the requisite concepts and formulations through which they may be understood; Section 3 focuses on neo-Hebbian formulations that employ more complex factors to guide exploration, with analysis of current state-of-the-art methods enhanced by consideration of results and theories thereon from neuroscience and psychology; Section 4 expands on the individual review analyses from Section 3 with concepts from the broader literature on learning theory to suggest methods by which their approaches to the exploration–exploitation balance may be extended; and Section 5 offers concluding remarks which emphasize key takeaways from the reviewed material.

1.1. Motivations and scope

Deep learning (DL) derived approaches to decision and planning problems have dominated many headlines in recent years, fueling fervor for machine learning research with superhuman performance on a number of challenging but narrow tasks — most readers will be aware of the progress in competitive games achieved by the AlphaGo Zero system (Silver, et al., 2017), for example. While these performance accomplishments are extraordinary and have added much to the study of computational optimization for RL, they do not represent new knowledge in our theoretical understanding of value-driven learning. Hebbian plasticity often serves as a set of models which allows for testable interplay between the experimental observations of biological neuroscience and the conceptual predictions of computational neuroscience. While this may seem a purely academic benefit, the interaction between these research domains has a long and fruitful history which has provided much of the theory underpinning today's top performing computational approaches (Gerstner et al., 2018; Kuriscak, Marsalek, Stroffek, & Toth, 2015).

This review focuses on the exploration–exploitation dilemma, a longstanding open problem in general RL (Sutton & Barto, 2017), in the context of neo-Hebbian RL. Computational RL tasks an agent to learn through trial and error interaction with a given environment. This is accomplished by sampling actions in visited states of its environment, observing the consequences of the actions taken in each state, and updating its action selection policy to maximize a cumulative reward value. The aim of maximizing the reward signal received through acting within the environment is known as exploitation. Given some experience acting within the environment, the agent is able to exploit the knowledge acquired about its reward structure with the expectation that pursuing the highest valued states (or state–action pairs) will maximize the reward received. For such an exploitative strategy to be effective at reward maximization, the agent must sufficiently sample the state–action space of its environment by exploring previously unvisited states as well as previously untried actions at visited states.

For most tasks of interest in computational RL, it is simply unfeasible to perform an exhaustive exploration of the environment and its responses to the agent's action space, as the state and action spaces may be quite vast and possibly continuous. Further, state transitions and reward functions may be stochastic, requiring many visits to calculate an accurate expectation of their value. These factors make brute force exploration strategies highly inefficient in the best case and practically indefinite in the worst. Given this abstract view of the trial and error learning challenge, we can define the exploration–exploitation dilemma in terms of the trade-off or balance that must be maintained between these competing aims during the course of learning to act within an environment. As the agent can neither always explore nor always exploit to efficiently learn a useful policy, some mechanism must be employed to determine when the agent should exploit previous learning and when it should instead explore available options that may or may not result in increased cumulative rewards.

The significance of our scope has practical bearing on the current state of RL research. In many tasks of interest both to theoreticians and engineers, the reward landscape of a defined environment is often sparse and/or non-static (Gregor & Spalek, 2014; He & Zhong, 2018; Hu, Song, & Huang, 2019; Machado, Bellemare, & Bowling, 2020). This means that the reward signals available may often fail to guide the learning algorithm towards optimal or near-optimal solutions. Handcrafting reward structures using domain knowledge can often allow experimenters to circumvent this problem without producing a solution to it. There are also numerous techniques that combine random or semi-random decisions with variations on the learning rate parameter in an effort to increase the likelihood of the model discovering informative structures in the environment by chance. Neither case represents a general solution to the open problem of plausibly stimulating exploratory behaviors in computational agents.

Neo-Hebbian RL provides a framework wherein additional factors beyond the reward signal may be incorporated to influence the rate and direction of learning. The interaction between extrinsic and intrinsic motivational signals in neo-Hebbian RL has been shown to produce naturalistic exploratory behavior using only simple models of these factors. We bring focus to these methods to assess them in light of more recent research from psychology and biological neuroscience, with the aim that their improvement may inform future approaches to the exploration–exploitation balance across the RL paradigm.

1.2. Related works

Hebbian plasticity is a well-established framework for algorithmically representing the physiological changes induced by coincident activity between connected neurons. As a newer approach to biologically plausible learning which combines purely local factors with global value signaling to enable innate RL in the model, neo-Hebbian research has sparked a number of insightful review articles. These include works addressing neo-Hebbian RL for both rate and spike encoded neuron models, and we briefly discuss the significance of these related references in this section.

Shouval et al. (2010) discussed R-STDP (the spiking class of neo-Hebbian plasticity formulations) in a broader context which considers the temporal interplay of observed biophysical changes during learning at various granularities (milliseconds, seconds, minutes, etc.). Drawing from results in biological neuroscience, the authors focused on experimental evidence suggesting that natural learning rules involve a high dimensional parameter space that may be better modeled by focusing on the dynamics of intracellular calcium messaging instead of spike trains. Under their perspective, STDP operates on a relatively fine temporal scale to induce intermediate changes such as the creation of synaptic tags. Temporally broader biophysical changes, such as the release of dopamine to portions of the brain in response to certain environmental conditions, then serve to regulate the learning rules which operate on the changes at finer granularities. While the debate between the significance of calcium messaging and that of neural spike events is beyond our present scope, this work provided well-reasoned arguments suggesting that the interplay between Hebbian (spike timing) and neo-Hebbian (neuromodulation) factors is more complex than can be represented by the gating and scaling effects provided by using only a simple signal representing extrinsic reward.

Feldman (2012), with a similarly biological emphasis, provided a broad overview of STDP focusing on the impact of spike timings relative to other factors such as firing rates, cooperative and competitive neural activation patterns, and so forth. Emphasis here was placed on the cellular mechanisms believed to implement STDP on a biological level, including the manner by which dopamine alters the shape of the STDP window at synapses involving different types of biological neurons. This work provided additional bolstering of neo-Hebbian theory through its analysis of experimental results illustrating variations in the effects of dopamine on inducing plasticity changes at the cellular level under varied conditions. This may be interpreted as indication that a more complex model of neuromodulation is needed to account for seemingly contradictory effects such as the synaptic conversion of long-term potentiation (LTP) to depression (LTD), among other phenomena.

In Tetzlaff et al. (2012), the topics of learning and memory are decomposed along a temporal axis to argue that STDP should be considered a prominent factor in a more complex learning framework. By dissecting the learning problem according to this temporal scale, the authors presented a solid evolutionary argument with strong support from current neuroscientific research for the existence of differing yet overlapping mechanisms of learning to coexist in highly developed neural systems. From this perspective, the emergence of reinforcement as one of the slowest and longest acting mechanisms that influences the learning of perception and behavior is a natural consequence of the scale and constraints inherent to the problems that biological RL addresses. Not only are real rewards sparse and highly dynamic but real neural networks, while undeniably powerful in their problem solving capacity, are evolved adaptations subject to metabolic limitations, finite channel capacity, propagation delays, and likely numerous more factors that impact the computations they perform. While this work did not focus exclusively on topics pertinent to RL, we found its suggestions recommending an increased research focus on the functionality of inhibition and learning at inhibitory synapses to be particularly apt.

Frémaux and Gerstner (2016) provided one of the first comprehensive reviews to focus exclusively on R-STDP methods — many of these spiking neo-Hebbian approaches had been previously developed by the authors of this paper. In their overview of spiking three-factor STDP RL formulations, the authors focused on neuroscientific evidence for gating versus multiplicative impacts to STDP induction under modulation by dopaminergic reward. The authors further discussed how to accurately model the quantity of dopamine both during and between rewarding events. Although the available evidence discussed in relation to their focal topics is largely inconclusive, this work is an excellent entry point for an accessible but in-depth introduction to R-STDP methods.

Kuśmierz et al. (2017) offered a more topical treatment of multi-factor Hebbian learning which was not limited to the domain of RL. This perspective considered a number of potential factors beyond the typical use of dopamine reward signals, including other neuromodulators like acetycholine or serotonin. This expanded focus drew attention to the potential roles that such additional factors may serve during the learning process, and it was speculated that the combination of several additional factors may enable the coexistence of different types of learning in a single biological system by providing distinct value and error signals at differing temporal and spatial granularities.

Roelfsema and Holtmaat (2018) focused on the topic of synaptic plasticity in sensory cortices. This work integrated a wealth of experimental results to assess the plausibility of neo-Hebbian theory, covering support for (and in some cases against) theories relating to the functionality of synaptic tagging, the sharpening of receptive fields under directed attention, among other pertinent open issues in this domain. Their coverage included substantial analysis of the potential role of feedback connectivity involving inhibitory interneurons in learning during sensory processing.

Gerstner et al. (2018) devised a categorical system for assessing various neo-Hebbian RL approaches in the literature and analyzed each prospective sub-class in the context of the available support for their existence in biological nervous systems. This paper provided discussion on the potential for surprise-driven learning to occur in neo-Hebbian contexts, although their definition of Hebbian learning was arguably stretched by including sub-threshold potentials in addition to spiking action potentials. The authors additionally argued that separate eligibility traces for potentiation and depression at the synapse should be maintained. These traces would follow non-identical dynamics to stabilize dynamics in recurrent circuits and potentially permit the emergence of natural event prediction mechanics, a perspective which we find well supported by both the neuroscientific and computational literature.

2. Background

To enable a more rigorous treatment of the topics of neuromodulation and RL under neo-Hebbian plasticity, this section briefly introduces the standard notation for the computational models discussed in later sections, provides pertinent background detail on some concepts of significance to the topic of learning, and introduces the prevailing paradigms of artificial learning algorithms.

In the following sections, when discussing the dynamics for a given pair of neurons connected via synapse, we assign the pre-synaptic unit the index j and the post-synaptic neuron the index i. We may represent the output or "activation" of each unit by y_j and y_i, respectively. These activation values, which may be either continuous or discrete, will typically refer to the output of some (potentially non-linear) weighted function of the unit's inputs. The weights used for the calculation of a neuron's activation are the target of the learning or plasticity rules discussed herein, and we represent the general efficacy of the pre-synaptic unit j in affecting the activity of the post-synaptic unit i by the weight $w_{j,i}$. For a thorough introduction to the modeling of neuronal dynamics, including the wide variety of neuron models found in the texts we discuss later, we refer interested readers to the standard text found in Gerstner, Kistler, Naud, and Paninski (2014).

In the remainder of this review, we refer to computational models of neurons as being either spike-based or rate-based. This distinction is most easily understood from the perspective of the temporal granularity used in the modeling of neural activity. When we refer to spike-based neural network models, this signifies a finer temporal granularity of the neuron model that captures the timing $t^{(f)}$ of fired action potentials. This granularity contributes an increased biological realism to SNN simulations, capturing the temporal relationship whereby spikes from pre-synaptic neurons drive changes to the membrane potential of post-synaptic units that then may or may not cause the post-synaptic unit to fire at a later time.

Maintaining a record of the timing of spikes for both pre- and post-synaptic neurons allows for learning rules that consider both the temporal distance and order of spike events as factors, which we discuss in greater detail in Section 2.1.3. Although the focus of this review is limited to neo-Hebbian learning rules and the mechanics by which they manage the exploration–exploitation balance in RL tasks, there are a several prominent spike-based neuron models that underpin implementations of these neo-Hebbian networks. These include frequently used models in the computational neuroscience literature such as Integrate-and-Fire (IF) models (Lapique, 1907; Tuckwell, 1988), which are relatively expedient computationally and have been well studied in a number of model variations (Fourcaud-Trocmé, Hansel, van Vreeswijk, & Brunel, 2003; Hansel & Mato, 2001; Latham, Richmond, Nelson, & Nirenberg, 2000), as well as neuron models that more faithfully reproduce biological neural data such as the Izhikevich model (Izhikevich, 2003) and the Spike Response Model (SRM) (Gerstner, 1990; Gerstner, Ritz, & Van Hemmen, 1993). Paugam-Moisy and Bohte (2012) provide a highly accessible introductory treatment of these commonly employed neuron models.

Rate-based neurons model activity at a coarser temporal granularity than their spike-based counterparts, condensing the timing details of individual spikes into an average rate over uniform windows of time. Typically, simulations for rate-based ANNs do not actually model the spiking activity of biological neurons or the changes in membrane potential associated with it. Rather, rate-based neurons produce activation or output values as a (in most modern cases, non-linear Apicella, Donnarumma, Isgrò, and Prevete (2021)) function over a weighted sum of their pre-synaptic inputs. These activation values then represent the average spike count the unit would produce as output (to any post-synaptic neurons) over the next temporal window, although this aspect of biological realism is often neglected to allow for negative activation values that are understood to be inhibitory as negative rates over time are not biologically possible.

2.1. Adaptation and synaptic plasticity

Plasticity rules are attempts at computationally recreating the persistent changes in efficacy observed at synaptic junctions between pairs of neurons. These rules form the basis of learning algorithms for the neural network models in the context under review. As neo-Hebbian RL significantly extends Hebbian learning, in this section we provide an introductory treatment of the concepts and prominent formulations for this foundational class of bio-plausible learning algorithms.

2.1.1. Hebbian learning rules

Often cited and occasionally poorly paraphrased, the modern foundation of correlation-derived learning, Hebb's postulate (famously stated on page 62 of Hebb (1949)), proposes: "when an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased". The concept Hebb described is now referenced as long-term potentiation (LTP) at the synaptic level, through which the relative efficacy of the connection between two neurons, typically treated as a weighting value of the connecting synapse, increases due to the effects of a pre-synaptic action potential contributing to the generation of a post-synaptic pulse.

We now know that correlated LTP alone is insufficient to capture the complex synaptic dynamics observed experimentally, so additional formulations for observed long-term depression (LTD) of the efficacy, or weight, of a synaptic connection are required (Dong, et al., 2012; Malenka & Bear, 2004). These may take the form of anti-Hebbian learning formulas, wherein pre-synaptic pulses driving post-synaptic spikes result in a net reduction of the impact of the synapse in question, or of frameworks that model the impact of improperly ordered spike pairs (a reversed spike ordering from standard Hebbian description). Other plausible mechanisms consider the concept of atrophy upon the synaptic connection as a factor that counteracts the monotone increases dictated by correlated LTP under Hebbian learning.

In the context of rate-based Hebbian learning, simultaneous correlated neural activation is typically (but not always) employed rather than an explicitly causal model which incorporates the timing of neural activations. If the rate of activation of two connected neurons are both coincidentally heightened/dampened, we may consider their activity to be correlated; if the activation rates diverge between the same pair of units, we consider their activity to be anticorrelated. These cases are typically associated as conditions to trigger LTP and LTD, respectively.

$$\tau_w \frac{dw_{j,i}}{dt} = f(w_{j,i})(y_j - y^{out})(y_i - y^{in}) \tag{1}$$

Eq. (1), adapted from Kuriscak et al. (2015), provides a stable form of rate-based Hebbian learning which models both LTP and LTD. Weight changes for the synapse connecting unit j to neuron i occur as a product of three expressions: (i) a function on the current weighting of the connection, which may introduce non-linear dynamics to weight changes; (ii) the difference between the current output firing rate of the post-synaptic neuron i and an upper bound threshold for output firing rates y^{out}; and (iii) the difference between the current input firing rate from pre-synaptic unit j and the respective upper bound threshold for input firing rates y^{in}. The use of differences between firing rates and threshold parameters, which can allow one or both of the latter multiplicative factors to take a negative sign, enables LTD to counteract unstable growth of the weight. This particular formulation implements both pre and post-synaptic gating when both threshold parameters are non-zero.

$$\Delta w_{j,i} = \alpha y_i(y_j - y_i w_{j,i}) \tag{2}$$

Oja's Hebbian learning rule, shown in Eq. (2), allows for a given neural unit to perform principle component analysis over its inputs (Oja, 1982) and is among the better studied variations on rate-based unsupervised Hebbian learning. The expression in parenthesis is considered the effective input to the neuron and serves to stabilize the growth of weights against divergence. Parameter α serves as a learning rate to scale the magnitude of weight updates.

While many other variations on the basic Hebbian learning rule exist, these rules operate on the same essential components – the current synaptic weight and the activation rates of both pre- and post-synaptic units. It is also possible to model causality – the effect of the pre-synaptic unit on the post-synaptic one – by using activation values which are not coincident in time. An example of this class of rate-based Hebbian learning is Rarely Correlating Hebbian Plasticity (RCHP), introduced in Soltoggio and Steil (2013). The original form of RCHP is expressed below in Eq. (3).

$$\Delta w_{j,i} = \begin{cases} +0.5 \text{ if } y_j(t-\Delta t)y_i(t) > \theta^+ \\ -1 \text{ if } y_j(t-\Delta t)y_i(t) < \theta^- \\ +0 \text{ otherwise} \end{cases} \tag{3}$$

This model of rate-based Hebbian learning induces specified, asymmetric weight adjustments for LTP and LTD when the product of post-synaptic activity (at the current time) with pre-synaptic input (over a small window of time Δt) exceeds (LTP) or fails to meet (LTD) established thresholds for their magnitude – θ^+ and θ^-, respectively. By ignoring common correlated activity between units – that which falls between the lower and upper thresholds specified – this formulation allows for RCHP to adjust the weighting only for near-coincidental neural activations that are highly likely to be correlated (or anticorrelated in the case of LTD). This rule can produce similar learning of weights to STDP in a highly efficient manner, as rate-based neurons are more efficient to simulate.

2.1.2. Differential hebbian learning rules

While the rate-based form of Hebbian learning described above implements synaptic plasticity changes on the basis of coincident pre- and post-synaptic activations, differential Hebbian learning (DHL) rules compute updates to synaptic weights using rates of change (derivatives) in neuron activations. This enables the learning rule to account not only for correlation but also for causation in the propagation of neural signals. We can assess this in Eq. (4), which is derived from the original formulation for this class of Hebbian learning in Kosko (1986).

$$\tau_w \Delta w_{j,i} = \dot{y}_i \dot{y}_j \tag{4}$$

In the equation above, $\dot{y}$ refers to the derivatives of the pre- (j) and post-synaptic (i) activities. This requires a differentiable model of neural activity that captures transient increases and decreases, which can be obtained for discrete event models by using an appropriate kernel on the activations. The use of such kernels, where necessary, ensures that a monotonic increase in the activation of a given unit temporally precedes a peak activation that is followed by a monotonic decrease. Using the overlap in these transients of neural activity enables a temporally symmetric modeling of LTP and LTD phenomenon. Considering the case when pre-synaptic neuron j begins to increase in activation (signifying that $\dot{y}_j > 0$) shortly before unit i does the same, $\dot{y}$ is positive for both pre- and post-synaptic units, yielding LTP of the synaptic connection during their overlapping transient increases. This is followed by a brief period of LTD when $\dot{y}_j$ becomes negative before $\dot{y}_i$ does. If we reverse the ordering of these transient increases, the cumulative change to synaptic weighting remains the same as both $\dot{y}_i$ and $\dot{y}_j$ are positive until, in this case, $\dot{y}_i$ becomes negative first. The net effect in both cases is LTP or increased weighting. Conversely, when either unit begins a transient increase in activity while the other is expressing a transient decrease, their overlap yields opposing signs and thus produces LTD.

To better reflect the causal relationship between pre- and post-synaptic activities, Porr and Wörgötter (2003) introduced a temporally asymmetric variant of DHL called isotropic sequence order (ISO) learning. This DHL variant models causality by replacing the derivative of the activation of the pre-synaptic unit $\dot{y}_j$ with the current activation, yielding Eq. (5).

$$\tau_w \Delta w_{j,i} = \dot{y}_i y_j \tag{5}$$

This formulation by Porr and Wörgötter (2003) captures the same correlations in synaptic activity while enforcing the temporal ordering required to infer causality in pre-synaptic activity driving post-synaptic responses. If pre-synaptic neuron j is active while unit i is becoming more active ($\dot{y}_i > 0$), we can infer that the activity of j is at least partially driving the increase in activity expressed by i and the ISO rule produces LTP. When j is active (but not necessarily becoming more active) while unit i is becoming less active ($\dot{y}_i < 0$), we know that the activity of j is not driving this change in the activity of i and the rule induces LTD as a consequence. Similarly, dampened activations by unit j when i is experiencing a transient increase yields LTD while the same lack of activity by j results in LTP when i is becoming less active.

Zappacosta, Mannella, Mirolli, and Baldassarre (2018) constructed a framework to unify the variety of first-order DHL rules proposed in the literature, yielding general differential Hebbian learning or G-DHL. Their formulation considers eight components, divided evenly into differential (both factors being derivatives as in Eq. (4)) and mixed (derivative and non-derivative as in Eq. (5)) additive factors. The total of eight factors is derived by decomposing the derivatives of pre- and post-synaptic units into their positive and negative components. Each of these factors can be manipulated via hyperparameter to influence their inclusion/exclusion (non-zero or zero), their direction of influence (LTP or LTD based on sign), and their contribution to weight updates (their magnitude). The flexibility afforded by this generalization of DHL enables the reproduction of many experimentally observed neural phenomenon, as demonstrated in Zappacosta et al. (2018), and it may be employed for both rate and spike-coded neural models.

2.1.3. Spike-timing-dependent plasticity

STDP provides a framework for formulations of biological and artificial spiking neural systems consistent with the causal relationship central to Hebbian learning and compatible, by extension or modification, with anti-Hebbian phenomena. As the name implies, STDP rules employ the timings of spikes (in the simplest case pairs, as treated here) to determine the appropriate change in synaptic strength between the units involved. We present here the formulation for a basic pair-based STDP update rule. Consider a pair of neurons, j and i, connected as pre-synaptic and post-synaptic units, respectively. We denote the times of the events of their pre-synaptic and post-synaptic action potentials as t_{pre} and t_{post}, defining a measure on their temporal separation as $|\Delta t| = |t_{post} - t_{pre}|$. A simple update rule for the weight $w_{j,i}$, adapted from Gerstner et al. (2014), is expressed in Eq. (6).

$$\Delta w_{j,i} = \begin{cases} A_+(w_{j,i})e^{\frac{-|\Delta t|}{\tau_+}} & \text{at } t = t_{post} \text{ for } \Delta t > 0 \\ A_-(w_{j,i})e^{\frac{-|\Delta t|}{\tau_-}} & \text{at } t = t_{pre} \text{ for } \Delta t < 0 \end{cases} \qquad (6)$$

This formulation permits both flexible Hebbian and anti-Hebbian dynamics through the selection of adaptation functions, $A_+(w_{j,i})$ and $A_-(w_{j,i})$, which may model alterations of efficacy as a function of the current synaptic weighting. The decaying exponential term, including the LTP and LTD decay constants τ_+ and τ_-, reflect the principle of temporal locality in STDP update rules; spike pairs relatively close in time (typically on the order of tens of milliseconds) are less coincidental and experimentally evoke stronger adaptive responses than more remote pairs.

This principle can be appreciated visually by assessing Fig. 1, which graphs the magnitude of plasticity changes with respect to the timing difference between pre and post-synaptic spiking for LTP and LTD. We refer to the distinction between LTP caused by a pre-before-post spike pair ordering and LTD arising from a post-before-pre spike pairing as causal and acausal forms of STDP, respectively. In the case of LTP under STDP, a pre-synaptic action potential contributing to the generation of a spike at the post-synaptic neuron shortly thereafter implies a causal relationship under Hebb's postulate. Pre-synaptic activity that follows a post-synaptic spike is inherently acausal, having not served to drive the post-synaptic pulse which preceded it. Markram et al. (2011) contain a highly accessible overview of the concepts and history behind the STDP framework that provides a deeper intuition on the biological considerations behind its development.

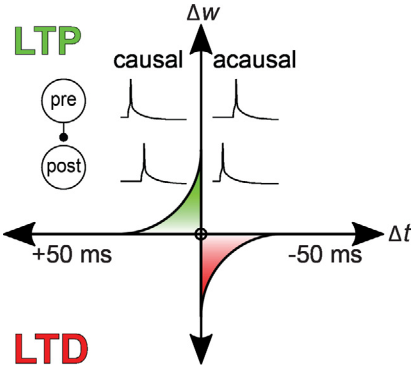

Fig. 1. Illustration of Δw (y-axis) with respect to Δt (x-axis) under a general STDP framework for both pre-then-post (causal, LTP, green) and post-then-pre synaptic pairings (acausal, LTD, red). Adapted from Markram, Gerstner, and Sjöström (2011). Note that the horizontal axis, corresponding to Δt, is reversed from the conventional left-to-right increase in value; the same is true in the originating source graph. As the magnitude of Δt increases between spike pairings, the corresponding synaptic changes to Δw become smaller — these spike pairings, pre-then-post on the left and post-then-pre on the right, are understood to be less indicative of a causal or acausal relationship between the spike pairing. Conversely, as Δt approaches 0 on the graph, indicating a shorter temporal interval between pre- and post-synaptic spikes, the STDP framework induces significantly stronger LTP (left, green) or LTD (right, red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Learning paradigms

In regard to computational learning theory, there are three classic paradigms: unsupervised, supervised, and reinforcement. These broad theories on the nature of learning may also overlap in some frameworks. This section briefly treats on the distinctions of these approaches and their better-known methodologies.

2.2.1. Unsupervised

Having already introduced the concept of Hebbian learning, whereby connections are strengthened through correlated activity, the fundamental functional characteristic of unsupervised learning is both simple and yet foundational to more advanced schemes for learning. Unsupervised learning mechanisms such as Hebb's rule (Gerstner et al., 2014) for rate-encoded neural networks or STDP in spike-encoded variants operate on the principle of strengthening or weakening synaptic connections irrespective of consequent network activity. This type of learning does not incorporate any measure of correctness or utility, yet the identification of associations, whether simply correlational or indicative of causality, continues to serve as an underlying factor for more complex learning capabilities.

From an evolutionary perspective, the unsupervised aspect of neural learning logically precedes any capacity for feedback — the ability to distinguish similarities and differences in stimulation must be possessed prior to the development of stimulus–response dynamics, for example. Given this view of a natural progression in the development of learned reactions predicated on the precedence of unsupervised mechanics, we treat the ensuing learning paradigms as mechanisms implementing selectivity, among more intriguing dynamics, atop this foundation.

2.2.2. Supervised

Supervised learning methods incorporate the concept of correctness to induce selective responses; this requires an explicit error signaling mechanism in addition to a determined source of ground truth. The most prolific format of supervised training common in the literature is the backpropagation (BP) method (Rumelhart, Hinton, & Williams, 1986), famed for its deep learning successes in combination with the availability of large labeled datasets and efficient gradient of error calculations (Shrestha & Mahmood, 2019). Deep learning methods have achieved remarkable success across many domains of machine learning, including models for image classification with deep convolutional networks such as GoogLeNet (Szegedy, et al., 2015) and language translation with attention mechanics such as the Transformer architecture (Vaswani, et al., 2017). While other supervised learning approaches exist (Lee, Zhang, Fischer, & Bengio, 2015; Wang, Belatreche, Maguire, & McGinnity, 2014; Zenke & Ganguli, 2018), BP is the predominant form of supervised training for artificial learning algorithms and perhaps the most exemplary of the paradigm.

A more interesting and potentially pertinent variation on the concept of supervised learning is a self-supervised approach. While this approach has received more attention for its use in static generative modeling, such as in autoencoders (Hinton & Salakhutdinov, 2006) and generative adversarial networks (Goodfellow, et al., 2014), the use of these methods with recurrent networks can produce powerful sequential prediction models (Jawed, Grabocka, & Schmidt-Thieme, 2020). In these methods, the ground truth signal is not a handcrafted indicator of correctness (such as labels for classification tasks) but rather a withheld or otherwise unseen (to the learning agent) portion of the training data. For time series data, this requires the model to produce output intended to predict the next value, corresponding to some $t + \Delta t$, after having received inputs over the series up to the stimulation corresponding to the current time t. The predicted next stimulus may then be compared with the actual next input data from the series and many error correction-based learning algorithms, such as BP methods, may be used to improve the predictive prowess of the agent. This approach to learning over sequential data can be related to certain predictive or planning methods for RL, as seen in Pathak, Agrawal, Efros, and Darrell (2017) for example.

2.2.3. Reinforcement

As this work focuses on reinforcement learning (RL) with respect to bio-plausible learning models, we present here a brief introduction to the classic reinforcement learning formulations from the broader domain of machine learning, adapted from the modern text by Sutton and Barto (2017). In a neo-Hebbian context, RL builds upon the unguided coincidence detection of unsupervised learning by incorporating the concepts of reward and punishment, rather than the extension with evaluation by predetermined correctness employed by supervised learning methods.

The concept of reinforcement of learned behaviors is well established in the study of operant conditioning, whereby the voluntary response of an organism to a given stimulus is modulated either to increase or decrease the probability of that response in the future. This conditioning occurs through repeated observance of net positive (reinforcing, either through exposure to a pleasant stimulus or removal of an aversive one) or net negative (punitive, either through subjection to an aversive stimulus or elimination of a pleasant one) outcomes. In conjunction with early results from the study of dynamic programming, computational RL theory sought to develop learning agents capable of adapting to experiential feedback inherent to a defined environment, rather than through instruction by an explicit error signal. Accomplishing this form of learning requires an agent to both explore its environment and to learn to exploit the information gleaned from its interactions with the environment to most effectively maximize (minimize) the cumulative reward (punishment) to which it is subject over some typically variable temporal scale.

Temporal-difference (TD) methods are a central framework in the domain of RL. The simplest TD formulation, known as TD(0) for its one-step temporal window, calculates an update to the internal estimate of an environmental state's value $V(S_t)$ following observance of $V(S_{t+1})$ and any reward or punishment R_{t+1} generated due to the activity of the agent at the former state S_t; more abstractly, the underlying expectation of the value of the former state is updated by the experience obtained from interacting with that state through some available form of activity.

$$V_{t+1}(S_t) = V_t(S_t) + \alpha[R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)] \tag{7}$$

The one-step state-value update rule for TD(0) is given in Eq. (7), where α is a learning rate parameter, γ is a discounting factor accounting for the delay in obtaining the reward value of the successor state of the environment, and the term within the brackets is typically referenced as the TD error. These updates apply directly to the value function, an estimate of the actual long-term value of environmental states typically calculated by repeatedly sampling the environment through exploration of the state–action space. These value function updates impact the action selection policy through their role in determining the expected values of potential successor states. Note that an action with a potentially very high successor state value for a highly improbable state transition may not be selected by a greedy policy if a lower valued but more likely successor state is expected to offer higher cumulative rewards after taking a different action.

The one-step update of TD(0) can be generalized to account for a longer temporal window of experiential information on the value of a given environmental state by following the formulation of the TD(n) state value update rule, wherein the value of the state observed at time t is updated at time $t + n$ with a series of discounted returns generated through that n-step temporal window; the TD(n) state value update function is given in Eq. (8), noting that Eq. (7) corresponds to a reduction to $n = 1$.

$$\begin{aligned} V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha[\sum_{i=1}^{n}(\gamma^{i-1}R_{t+i}) \\ + \gamma^{n}V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t)] \end{aligned} \tag{8}$$

While the TD(n) method allows for effective value estimation using delayed rewards, the temporal duration of n is fixed. The TD(λ) method extends this approach by introducing an eligibility trace, denoted as λ, which allows for bootstrapping of rewards received in arbitrarily distant future states into the value estimate. The TD(λ) method is also more readily extensible to continuous-time modeling frameworks. Note that while Sutton and Barto (2017) use the symbol λ to represent multiple mathematical abstractions in various contexts throughout that text, we use λ exclusively to refer to an eligibility trace as is the standard in most recent works on neo-Hebbian RL.

Under the TD(λ) method, the eligibility trace λ for each state is used as a multiplicative factor on the TD error in the value update rule (Eq. (10)). The trace's value is a non-negative scalar which records how often a state has been visited and how recently these visits have occurred. The intuition behind eligibility tracing is that frequently visited states that precede rewards are more important for learning and their relevance has a temporal shelf life.

$$\lambda_t(s) = \gamma\lambda_t(s) + \boldsymbol{I}(S_t = s) \tag{9}$$

$$V_{t+1}(S_t) = V_t(S_t) + \alpha\lambda_t(S_t)[R_{t+1} + \gamma V_{t+1}(S_{t+1}) - V_t(S_t)] \tag{10}$$

The trace update in Eq. (9) captures this. The second term on the righthand side counts the number of times that a state has been visited using an indicator function $\boldsymbol{I}$ which takes the value 1 when the argument is true and 0 elsewhere. The first term causes the trace value to decay asymptotically to zero over time according to the trace decay parameter $\gamma \in [0, 1]$. As such, the value of λ for a given state reflects a function of both that state's visitation and its temporal relationship with delayed rewards, which is implicitly recorded by the amount of decay. Eq. (9) is adapted from Equation 7.5 in the first edition (Sutton & Barto, 1998) of the primary text used for this section, replacing a conditional equation with an equivalent indicator function. This particular formulation of eligibility tracing for computational RL was used as inspiration for synaptic eligibility tracing methods which can enable neo-Hebbian RL with distal rewards, as introduced in Section 2.3.2.

2.3. Modulation and reinforcement learning

We have given the relevant background for: (i) rate and spike-based neuron models, (ii) formulations for the basic Hebbian learning rules, and (iii) the main learning paradigms that can enable the extension of Hebbian learning rules. This section introduces a subset of the class of RL formulations commonly referred to as three-factor, or neo-Hebbian, learning rules. Neo-Hebbian mechanisms modulate (up or down) the change in strength between pre- and post-synaptic synapses, normally caused by a two-factor Hebbian rule, by incorporating the notion of value (or reward) as a third factor.

In its most basic form, neo-Hebbian RL alters standard Hebbian plasticity with various forms of scaling or gating in response to global reward signaling. The neuromodulator dopamine is often proposed to serve as this rapid signaling mechanism for reward in theories of behavioral learning in animals. This interpretation of dopaminergic neural activity has inspired a number of frameworks aiming to integrate the concept of TD reward errors with Hebbian learning theory (Frémaux & Gerstner, 2016; Gerstner et al., 2018). While alternative theories on the role of dopamine in learning have been proposed (discussed in later sections), a global reward signal is essential for extending Hebbian plasticity into more complex RL frameworks and its role must be understood before considering alternative modulating dynamics.

2.3.1. Hedonism and delayed reward

As a precursor to neo-Hebbian RL, Seung (2003) proposed "hedonistic" synapses modeled as stochastic processes modulated by a global reward signal (inspired by the study of dopamine neuromodulation). The spiking IF neurons employed in this framework encapsulated the concept of synaptic weighting into a probabilistic synaptic spike transmission formulation wherein learning was a relation on global reward and the probability ($p_{j,i}$) of a pre-synaptic (unit j) action potential generating the release of some amount of neurotransmitter to the post-synaptic (unit i) membrane neuroreceptors across the synapse; for the purposes of the model, this is viewed as successful spiking.

$$p_{j,i} = \frac{1}{1 + e^{-w_{j,i} - c_j}} \tag{11}$$

Eq. (11) formulates the probability of a pre-synaptic action potential successfully affecting the post-synaptic neuron as a logistic sigmoid function (having range (0, 1)) on the learned weighting (w) of that connecting synapse and the calcium concentration within the pre-synaptic neuron. The variable c represents a simple (and interchangeable) model of calcium dynamics for the pre-synaptic unit, increasing by a parameter Δc at the moment of pre-synaptic spike generation (regardless of whether that spike event triggers release of neurotransmitter to the synapse) and exponentially decaying by $dc/dt = -c/\tau_c$ thereafter. Plasticity for the weight component under this framework (Eq. (12)) was modulated by the global dopaminergic reward signal ($R(t)$) and an eligibility trace (Eq. (13)) which decays following $d\lambda/dt = -\lambda/\tau_\lambda$. Parameter η functions as a learning rate to scale weight updates.

$$\frac{dw_{j,i}}{dt} = \eta R(t)\lambda_{j,i}(t) \tag{12}$$

$$\Delta\lambda_{j,i} = \begin{cases} (1 - p_{j,i}) \text{ if spike neurotransmitter release succeeds} \\ -p_{j,i} \text{ if release fails} \end{cases} \tag{13}$$

The additive update rule applied to $\lambda_{j,i}$ during pre-synaptic spiking incorporates some dynamics of the eligibility trace used in TD(λ) (Eq. (9)), increasing with the accumulation of recent relevant activity and decaying with temporal distance. Seung (2003) further conceptualized this framework as a greedy approximation of gradient ascent through the parameter space of w, deviating from the additive indicator function employed in the TD(λ) to restrict the eligibility trace to have zero mean so as to prevent bias in the weight traversal of that search space. A biologically plausible implication (from an operant conditioning perspective) of their formulation is the following consequence to plasticity with respect to rewards: recent and successful spike propagations relative to a positive reward signal result in LTP at the synapse and successful spike propagations followed by a negative reward induce LTD. Conversely, firing failures preceding a positive reward result in LTD while the same failures before negative rewards induce LTP. This is in contrast to the eligibility tracing of Eq. (9), which is strictly non-negative and would lead only to LTP given positive rewards and only to LTD under negative ones.

2.3.2. Distal rewards and credit assignment

Inspired by the formulations laid out in Seung (2003), Izhikevich (2007) incorporated dopaminergic reward directly into a modulated R-STDP learning strategy. The idea involved modulating the effects of STDP (LTP and LTD) on weight updates using a function of both direct environmental reward and the impact of those environmental reward signals on the changing concentration of dopamine over time (rather than a single, direct reward model as employed in Equation 12 by Seung (2003)).

This work introduced a more complex eligibility trace formulation, given in Eq. (14) where $STDP(\cdot)$ corresponds to an STDP plasticity rule like that given by Eq. (6). The eligibility trace is multiplied by the received temporally decaying reward signal $R(t)$, yielding Eq. (15) where $\frac{dR}{dt} = -\frac{R(t)}{\tau_{\mathrm{DA}}} + \mathrm{DA}(t)$ and $\mathrm{DA}(t)$ is some function describing the dynamics of dopaminergic concentration. One example function modeling diffusive dopamine concentration dynamics used by Izhikevich (2007) was $\mathrm{DA}(t) = 0.5R(t - t_{\mathrm{R}})$ where t_{R} is the moment of receipt for the most recent reward value.

$$\frac{d\lambda_{j,i}}{dt} = -\frac{\lambda_{j,i}}{\tau_\lambda} + STDP(t_{post} - t_{pre})\delta(t - t^{(f)}) \tag{14}$$

$$\frac{dw_{j,i}}{dt} = \lambda_{j,i}(t) \cdot R(t) \tag{15}$$

Using their namesake Izhikevich spiking neuron model, Izhikevich (2007) employed the reward modulatory signal in conjunction with eligibility tracing to solve the distal reward credit assignment problem – the determination of assigning appropriate credit to synaptic weights with respect to their contribution towards rewarding or punitive performance over temporal scales – with spiking neurons in a framework consistent with STDP. The

Table 1
Summary of the formulations for R-STDP weight updates and eligibility trace calculations provided for comparison of the methods surveyed above. Note that $\Delta(t^{(f)}_{pre/post})$ corresponds to $t_{post} - t_{pre}$, the input term for spike pair-based STDP.

Source	Update rules
Potjans, Diesmann, and Morrison (2011)	$\Delta w_{j,i} = \alpha[\lambda_i \varepsilon_i][R(t) - R_{base}]$ $\Delta\lambda_i = -\frac{1}{\tau_\lambda}(\lambda_i - \sum_{t_i^{(f)}}(\delta(t - t_i^{(f)})))$ $\Delta\varepsilon_i = \frac{\varepsilon_i}{\tau_\varepsilon} - \sum_{t_i^{(f)}}(\varepsilon_i - \delta(t - t^{(f)_i}))$
Yusoffa and Grüning (2012)	$\Delta w_{j,i} = [\alpha + R(t)]\lambda_{j,i}(t)$ $\Delta\lambda_{j,i} = STDP(\Delta(t^{(f)}_{pre/post}))$
Ozturk and Halliday (2016)	$\Delta w_{j,i} = \eta[\bar{R} - R(t)]\lambda_{j,i}$ $\Delta\lambda_{j,i} = \tau_\lambda(STDP(\Delta(t^{(f)}_{pre/post})) - \lambda_{j,i})$

update rule in Eq. (14) can be viewed as an STDP-scaled form of the eligibility trace update from Eq. (9), which performs credit assignment for TD(λ) methods. In the context of neo-Hebbian RL, this trace performs credit assignment not for visited environmental states (as done by the TD(λ) method) but for activity over synaptic connections which precede rewards generated by the environment.

In this neo-Hebbian form, the Dirac delta function serves as the event indicator, comparable to $\mathbf{I}$ in Eq. (9), which increases the value for a given synaptic trace $\lambda_{j,i}$ at time $t = t^{(f)}$, where $t^{(f)}$ is the firing time of either the post-synaptic unit i (in the case of LTP) or the pre-synaptic unit j (for LTD), whichever occurs later in the spike pairing within the window for induction of STDP. This event-driven step-wise increase to the eligibility trace $\lambda_{j,i}$ is scaled by the value of $STDP(t_{post} - t_{pre})$, reflecting the dynamics of STDP with respect to the temporal difference between pre and post-synaptic spiking.

Many approaches to dopaminergic modulation of STDP since Izhikevich (2007) follow similar formulations, albeit adjusted to advance alternative aims beyond the credit assignment problem. For brevity in comparison of these spike-based neo-Hebbian works, we provide their formulations governing weight updates in Table 1.

Yusoffa and Grüning (2012) biased the effects of reward modulation on the eligibility trace using a learning rate α while training spiking units to associate delayed pairings of input stimuli, while Ozturk and Halliday (2016) achieved output spike train reconstruction by smoothing dopaminergic reward delivery with respect to average reward returns $\bar{R}$.

Potjans et al. (2011) formulated a more localized approximation of TD learning by modeling the reward signal $R(t)$ as fluctuations in dopamine concentrations relative to a baseline R_{base}, allowing the extension of an eligibility trace with an additional "activity" trace ε to reconstruct the TD error for each neural unit by interacting with the neuromodulatory dopamine concentration during weight updates.

Soltoggio and Steil (2013) showed that a rate-based equivalent of the R-STDP learning framework used in Izhikevich (2007) could achieve comparable results under a reward-modulated form of RCHP (see Section 2.1.1) in terms of learning in classical and operant conditioning tasks under delayed reward. Rather than using an explicit eligibility trace, as seen above in Eq. (14), Soltoggio and Steil (2013) deconstructed the synaptic weights to include long and short-term components such that $W_{j,i} = W^{lt}_{j,i} + W^{st}_{j,i}$. Changes to the short-term component of a given weight, $W^{st}_{j,i}$, which occur similarly to the updates of eligibility traces above, immediately impact the overall weighting of synaptic input at the post-synaptic neural unit but are not consolidated into the long-term weighting until the delivery of reward. This allows for the underlying Hebbian component of the three-factor neo-Hebbian formulation to perform unsupervised learning between rewarding events without inducing potentially erroneous permanent changes to the long-term weight.

$$\Delta W^{st}_{j,i}(t) = -\frac{W^{st}_{j,i}(t)}{\tau^{st}} + RCHP_{j,i}(t) \tag{16}$$

Eq. (16) illustrates the dynamics of neo-Hebbian RCHP on the short-term weight component, where τ^{st} governs the decay rate of short-term plasticity changes and $RCHP_{j,i}$ corresponds to Eq. (3). Consolidation of the total weight value for each synapse occurs at the moment of reward delivery such that the long-term weight changes according to $\Delta W^{lt}_{j,i} = R(t)W^{st}_{j,i}$. While the modulatory signal $R(t)$ responsible for induction of short-term plasticity, in Soltoggio and Steil (2013) an essentially immediately impactful eligibility trace, was modeled as a discrete event, this does not preclude extension to continuous-time modeling akin to the dynamics of dopamine used by Izhikevich (2007). An upper-bound threshold for dopamine concentration could be used to induce long-term LTP, with a complimentary mechanism for LTD applicable in experiments which require it. The reward prediction error theory of dopamine which inspired much of computational RL theory is based largely on the study of dopamine transients, phasic activity by dopaminergic neurons which significantly deviate the extracellular concentration of the neuromodulator above or below its tonic baseline quantity in response to valued stimulation.

While not explicitly focused on the trade-off between exploration and exploitation in RL, Soltoggio and Steil (2013) did briefly consider the potential impacts of their synaptic weighting split. Repeated rare correlated activity at the synapse can allow for the short-term weights to grow rapidly without necessarily impacting the long-term component, as these changes to short-term plasticity decay rapidly. This may allow for the network to explore more extreme portions of the weight space during learning episodes in a temporary fashion, with repeated reward receipt inducing longer changes which encourage exploitative strategies.

Soltoggio (2015) extended the rate-based neo-Hebbian RCHP framework of Soltoggio and Steil (2013) with a focus on the issue of catastrophic forgetting in continual learning experiments. Their approach conceptualized the factoring of synaptic strength into short and long-term components as an approximate mechanism for hypothesis testing, using the modulatory signal $R(t)$ as evidence for or against the likelihood of a reward following stimulus-action pairs. Their newer formulation, termed Hypothesis Testing Plasticity (HTP), eschewed modeling LTD as a consequence of anticorrelated neural activity (the rate-based approximation of acausal STDP in RCHP) in favor of a consistent but weak form of weight depression provided by a slightly negative baseline value of dopamine – a strong contrast to the positive baseline value used in both Izhikevich (2007) and Soltoggio and Steil (2013). This negative baseline value for the modulatory signal continually induces LTD in the short-term weight components, which then require more consistent associations between experienced reward outcomes and stimulus-action pairs to grow large. We view the negativity of this baseline concentration of dopamine as a computationally expedient mechanism for replicating otherwise biologically plausible weak LTD in the absence of reinforcement by reward despite the clear implausibility of a negative baseline value for any neuromodulator.

When combined with a threshold for induction, the second major deviation of HTP from neo-Hebbian RCHP which solidifies short-term plasticity into the long-term weight component upon any reward delivery, this formulation protects the stability of the network parameters in the long-term weighting by only adopting permanent changes which have accumulated substantial evidence through trial-and-error.

$$\Delta w^{st}_{j,i}(t) = -\frac{w^{st}_{j,i}(t)}{\tau^{st}} + M(t)RCHP_{j,i}(t) \tag{17}$$

$$\Delta M(t) = -\frac{M(t)}{\tau^{M}} + \alpha R(t) - b \tag{18}$$

$$\Delta w^{lt}_{j,i}(t) = \beta \mathrm{H}\left(w^{st}_{j,i}(t) - \Phi\right) \tag{19}$$

Eqs. (17)–(19) illustrate the distinctions between neo-Hebbian RCHP and HTP. Short-term weights are continually updated by rare correlated activity following the RCHP rule as before, but are now also continually modulated by the function $M(t)$ which models the extracellular dopamine concentration as a decaying function of received rewards relative to a negative baseline value $-b$. Long-term plasticity is additionally modeled on a continual basis using the heaviside step function $\mathrm{H}(\cdot)$, which takes the value $+1$ for positive arguments and 0 elsewhere; the threshold for long-term LTP, Φ, ensures that positive argument values only occur when the short-term weight exceeds the minimum for induction. β is a consolidation hyperparameter similar to a learning rate that governs the speed of induction into long-term weight changes. The authors included this parameter to model temporal delays in biological plasticity changes, though they noted that instantaneous induction ($\beta = 1$) gave similar results. To model long-term LTD changes, a symmetric match for Eq. (19) is simple to produce using only negation and an appropriate lower bound (Soltoggio, 2015).

2.3.3. Approximating the TD error

Q-learning, a family of TD algorithms focused on the optimization of value estimates for pairs of states and actions (Q-values $V_t(S_t, A_t)$ rather than the standard state values $V_t(S_t)$ related to Eqs. (7), (8), and (10)), is a staple of modern RL that addresses both the control (action selection) and evaluation (policy refinement) problems (Sutton & Barto, 2017). For problem domains where function approximation is necessary – typically those tasks involving a continuous rather than discrete set of states and actions – an Actor–Critic approach assigns the problems of control and evaluation to a pair of complimentary neural networks, dubbed "Actor" and "Critic" respectively, that work in an interleaved fashion to optimize the framework's approximation of the true Q-values for the task domain.

Frémaux, Sprekeler, and Gerstner (2013) extended the general framework of R-STDP introduced in Section 2.3.2 to follow this Actor–Critic network design with two networks of SRM$_0$ spiking neurons. Their formulation of the learning rule for both Actor and Critic neurons replaces the eligibility trace for temporal credit assignment with a smoothing kernel κ whose shape maintains an implicit and decaying record of causal (pre-before-post) paired spiking activity.

$$\Delta w_{j,i} = \alpha D(t)\left(\left[\mathbf{Y}_i(\mathbf{X}^{\hat{t}_i}_j \circ \varrho)\right] \circ \frac{\kappa}{\tau_R}\right)(t) \tag{20}$$

$$D(t) = \frac{R(t)}{N}\left[\sum_{i=1}^{N} \mathbf{Y}_i \circ \left(\kappa' - \frac{\kappa}{\tau_R}\right)(t)\right] - \frac{u_{rest}}{\tau_R} + R(t) \tag{21}$$

$$\kappa = \frac{e^{\frac{-t}{\tau_K}} - e^{\frac{-t}{\vartheta_k}}}{\tau_K - \vartheta_K} \tag{22}$$

Eqs. (20)–(22) describe the dynamics of weight updates for synapses connecting pre-synaptic units j to post-synaptic neurons i in terms of composite decaying learning rate α, TD error estimate $D(t)$, excitatory post-synaptic potentiation dynamics modeled by ϱ (formulation omitted for relevance), kernel κ (and its similarly omitted derivative κ'), kernel decay and rise temporal constants τ_K and ϑ_K, and pre- and post-synaptic spike trains $X^{\hat{t}_i}_j$ and Y_i of the form $\sum \delta(t - t^{(f)}_k)$ as in previous formulae. Note that the pre-synaptic spike train vector $X^{\hat{t}_i}_j$ is restricted only to spikes by pre-synaptic unit j that occurred prior to the most recent spike by post-synaptic neuron i, as signified by the superscript $\hat{t}_i$ which denotes the time of the last spike by i.

Neural units in both the Actor and Critic networks in the continuous control navigation tasks tested in Frémaux et al. (2013) received identical inputs from "place cells" whose spiking activity signals the location of the agent relative to the centers of discrete blocks in the state space. Both Actor and Critic spiking neurons employed the same weight update mechanism outlined above. Units within the Actor population received lateral connections between neurons indicating a preference to navigate in similar directions wherein each unit additionally potentiated those in its neighborhood of action preference and inhibited those whose spikes signal an incompatible choice. Combined with population vector coding, this scheme allowed a discrete number of neural units to encode continuous action choices via N-winner-takes-all action selection.

Frémaux et al. (2013) reported some encouraging experimental results, particularly in the learned behavior of Critic population neurons resembling that of biological dopaminergic "ramp" cells which have been observed to increasingly fire action potentials upon approach to an expected reward. This similarity to spiking activity in biological ramp cells is shown in Fig. 2. Further, their derivations illustrating the processes by which the TD error approximation is backpropagated for learning (despite being essentially undetectable within the observed spiking behavior of individual artificial neural units) hints toward the biological plausibility of some form of distributed backpropagation of value error under R-STDP methods that has, as yet, failed to be directly detected by neuroscientific research but is widely theorized to occur in biological reward-based learning under the reward-prediction error hypothesis.

3. Exploration: Beyond credit assignment

In the context of RL, exploration is the process by which an agent samples both the environment in which it operates and the available action space through which it may alter its relationship with that environment. This sampling process, when combined with a mechanism for estimating the long-term value of states (or state–action pairs in Q-learning), is used to enable exploitative action selection strategies for maximizing the quantity of reward received by the agent over time. This forms the conceptual basis for trial-and-error learning in computational RL theory.

While the value estimation performed by TD methods is highly efficient for dense and static reward landscapes under even very simple exploration strategies such as random search, the exploration process becomes a significant bottleneck in more complex environments. These are often not densely filled with reward-generating states and in some cases the reward generated in response to actions taken in a given state may change or disappear altogether. Learning exploitative action strategies in environments such as these then requires substantially increased sampling of the action and environment spaces, which has obvious impacts to both the temporal and computational efficiency of learning to perform tasks in these environments.

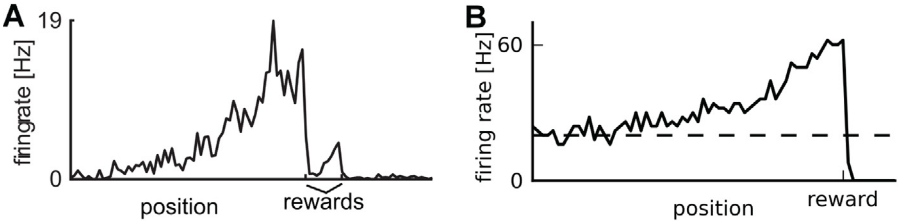

Fig. 2. Illustration of firing rate dynamics for A) rat ventral striatum "ramp cells" during a maze navigation task involving food rewards and B) single Critic population neuron during a linear track task.
Source: Adapted from Frémaux et al. (2013).

The reviewed material in prior sections has focused on the construction of modern neo-Hebbian RL methods which extend biologically plausible forms of Hebbian unsupervised learning to implement neural credit assignment through some mechanism of synaptic selectivity, ranging from tabular traces and short-term weights to more compact kernel methods. While a number of these works briefly consider the issue of stimulating exploration in neo-Hebbian RL agents, their methods for doing so are loosely equivalent to the semi-random search often used as a baseline in RL algorithms. The most common method used in general neo-Hebbian RL is to insert an additive noise function into the calculation of membrane potentials or weight updates to avoid consistently favoring exploitative action choices during learning (Huang, Wu, Yin, & Qiao, 2017).

This section considers works which have extended three-factor neo-Hebbian RL with the intention of developing exploratory behaviors in learning agents that are more capable of adapting to non-trivial reward landscapes, which may be sparse and dynamic.

3.1. Intrinsically motivated reinforcement learning

The concept of intrinsic motivation of behavior and its distinction from extrinsic motivators stems from the study of animal and human behavioral learning in (extrinsically) value-neutral settings (Baldassarre, et al., 2014). The prevailing hypothesis prior to the conceptualization of intrinsic motivators was that biological behavioral learning occurred following a drive to reduce some physical need relative to the survival and reproduction of an organism (Hull, 1943).

If we take physical hunger as an example of one such driver of behavioral learning, then we may view directed foraging as a behavior reinforced by consequent success in reducing the animal's need for nourishment under this theory of drive reduction. While this theory sufficed for experimental evidence around animal conditioning and goal-directed learning, it failed to account for much of the observed behaviors in situations lacking an obtainable goal or reward within the external environment of the organism.

The use of the terms "intrinsic reward" and "intrinsically-motivating" behavior was popularized in the literature on behavioral and developmental psychology by Harlow (Harlow, Harlow, & Meyer, 1950), among others, in attempt to explain behaviors through which organisms expend time and physical effort (finite resources from an evolutionary perspective) with no apparent goal or external benefit. Taking example from the study cited above, rhesus monkeys were shown to learn to efficiently solve non-trivial puzzles in a controlled environment without external incentives such as food. Further, the introduction of extrinsic rewards (treats) during the learning process was found to disrupt rather than enhance performance, leading Harlow to propose that a "manipulation drive" may account for the reinforcement of their puzzle-solving behaviors. Similar drives were proposed across the literature to account for a number of human and non-human behaviors associated with concepts such as curiosity, play, and investigation or exploration. These abstract drivers of behavior are called intrinsic as they lack any direct and discernible connection to an external goal or reward.

While the study of intrinsic motivation has a well-established history in the domain of psychology, its introduction to RL theory is more recent. In Schmidhuber (1991), the issue of efficient exploration by artificial agents led the researchers to propose an RL agent architecture with an adaptive world model. The intent of their design was to produce a system that could dynamically identify poorly modeled portions of the environment space which could then be targeted by a greedy exploration policy to expedite the learning process — essentially a form of meta-learning aimed at autonomously identifying where exploration would be most informative. This was accomplished by jointly training their world model with a "confidence module". This module was trained through supervised gradient methods to approximate the magnitude of weight changes induced in the world model during its training. These approximated improvement values were then used as reinforcement for maximization by the control module of their system. Through learning to estimate the performance improvements made in the world model via sampling of state space transitions, this confidence module was shown to significantly improve the efficiency of training the world model on a given environment.

This approach to incorporating intrinsic motivators in computational RL was formalized in Barto, Singh, Chentanez, et al. (2004) and Chentanez, Barto, and Singh (2004), which focused on hierarchical skill learning. Although their framework was based on option theory and utilized a predetermined measure of salience as additional reinforcement, their work served to illustrate the utility of intrinsic RL for acquiring complex action sequences to attain sparse extrinsic rewards. Schembri, Mirolli, and Baldassarre (2007) expanded upon the methods of Chentanez et al. (2004) both to generalize this approach to handle continuous state and action spaces as well as to autonomously generate additional intrinsic reinforcement signals. These signals were produced via a neural network trained by an evolutionary algorithm and supplemented a simple prediction error calculation to form a surprise-based intrinsic motivation value.

Singh, Lewis, Barto, and Sorg (2010) rigorously assessed the evolutionary significance of intrinsic motivation for RL in the context of optimal reward functions. Their experiments proved that the optimal reward function for a given agent, which may utilize intrinsic rewards to reinforce intermediary behaviors in addition to the primary extrinsic rewards, will outperform or at least equal the performance of the same agent that uses only a fitness-based reward function. Fitness, in this context, relates to the ability of a learning agent to achieve goals, such as successful hunting by wild animals or cumulative point acquisition in the context of a game.

The logic for their theoretical assessment of motivation in the context of evolutionary fitness is straightforward and bears weight on the use of intrinsic motivation in RL. An entirely fitness based reward function, corresponding to the case of extrinsic-only RL, only rewards behaviors illustrated via experience to confer fitness for the task at hand. Reward functions that additionally account for other factors, such as intrinsic motivators, reinforce intermediate behaviors that may enhance fitness at some later point for the agent — in addition to the same reinforcements provided by the fitness component of the reward function. This does not imply that every possible type and strength of intrinsic reinforcement will produce equal or improved performance compared to fitness-only reinforcement but rather that any given fitness function can be improved upon by balancing it against one or more other factors that incidentally enhance cumulative fitness. How to construct such an improved reward function in the general case remains an open problem in computational RL, and it is possible that successfully managing the exploration–exploitation balance requires a sufficiently good approximation of optimal reward functions.

3.2. Acetylcholine and R-STDP

Acetylcholine is thought to play a role in a number of neural functions, including the consolidation of memories (Fink, Murphy, Zochowski, & Booth, 2013; Golden, Rossa, & Olayinka, 2016), spatial learning (Zannone, Brzosko, Paulsen, & Clopath, 2018), and attention to unexpected changes in stimulation (Brzosko, Zannone, Schultz, Clopath, & Paulsen, 2017). While R-STDP has been successfully employed in spiking models on tasks with stationary targets such as supervised classification (Hao, Huang, Dong, & Xu, 2020) or spike train sequence reproduction (Ozturk & Halliday, 2016), applications of R-STDP methods to RL problem domains with spiking neuron models have inherited some issues from their TD learning foundations. These relate to the reward landscapes of realistic environments, which are often sparse in terms of non-zero reward values (Machado et al., 2020) and dynamic (Hu et al., 2019).

Learning from extrinsic reward alone in environments with sparse and/or dynamic rewards has proven quite challenging for diverse sets of model agents. Intrinsic rewards have been introduced as a compensatory mechanism to aid learning when the reward space is insufficiently informative to guide exploitation (Gregor & Spalek, 2014; He & Zhong, 2018). While the application of intrinsic reward methods has largely been a feature of the gradient-based deep RL approach, we present in this section a brief overview of recent efforts to incorporate some form of intrinsic modulation of R-STDP with spiking neurons.

The majority of works addressing the concept of cholinergic modulation of R-STDP in SNNs employs the modeling of acetylcholine as a complementary factor to counterbalance the influence of dopamine modulation on STDP. Dopamine modulation which follows the general form outlined in the previous sections results in learning which closely follows TD methods. This entails a complete bias in weight updates towards exploitative strategies, as reinforcement alone only solves the credit assignment problem but does not directly encourage exploration of the state and action spaces in general (Sutton & Barto, 2017).

The formulation in Golden et al. (2016) modeled the purported dynamics of acetylcholine as dampening LTP by imposing a linearly decaying form of the learning rate parameter η (see Eq. (12) for a corresponding constant learning rate equation); as such, their plasticity mechanism (a standard STDP formula like Eq. (6)), eligibility trace (Eq. (23)), and consequential weight update rule (similar to Eq. (12) but with an STDP eligibility update rather than a probabilistic formula) did not differ in any substantive way from the dopaminergic formulations presented in Section 2.3.2.

$$\Delta\lambda_{j,i} = -\lambda_{j,i} + \eta STDP(t_{post} - t_{pre}) \tag{23}$$

Each training trial would incur a small decrement to η which simplistically modeled the effect of reduced levels of acetylcholine due to repeated stimulus exposure. This monotonic decrease in the learning rate was intended to capture the loss of agent surprise when returning to previously visited states due to trial repetition, with the decaying learning rate serving to enforce smaller weight updates as training progressed. The cause behind the findings in Golden et al. (2016), where a combination of dopaminergic and cholinergic modulation reduced convergence of performance in comparison to a dopamine reward baseline framework (where learning rate η remains constant), should be mathematically apparent.

We turn now to more advanced attempts at combining dopaminergic reward with cholinergic modulation by addressing the group of efforts made toward applying sequential neuromodulatory mechanisms (compared to the direct acetylcholine modulation of dopamine modulation embodied in the methods of Golden et al. (2016)). Brzosko et al. (2017), extending their previous work showing that dopamine signaling served to lengthen the time window dynamics under STDP, sought to encourage exploratory behavior by combining acetylcholine with reward signaling in simulations of dynamic environments. This sequential approach employed an alternating (see Eq. (25)) formulation of the effects of neuromodulation, with acetylcholine driving LTD on active synapses over timescales with low dopaminergic reward and with dopamine inducing LTP over eligible timescales, including those corresponding to periods of high cholinergic concentrations, as consistent with previous neuronal studies.

$$\Delta w_{j,i} = \eta A(\sum_{t^{(f)}_{pre/post}} STDP(t_{post} - t_{pre}) \cdot \lambda_{j,i}) \tag{24}$$

$$\Delta A = \begin{cases} -1 \text{ for } DA^{-}, ACh^{+} \\ 1 \text{ for } DA^{+}, ACh^{+} or ACh^{-} \end{cases} \tag{25}$$

The framework provided by Brzosko et al. (2017) improved upon the form of acetylcholine modeling employed by Golden et al. (2016) by applying an alternating rather than monotonically decaying learning rate η, where $\eta = 0.002$ in the presence of acetylcholine without dopamine and $\eta = 0.01$ during dopaminergic signaling. Further, their equation for the temporal decay of the eligibility trace λ alternated in effect according to the presence of dopamine, capturing the purported dynamics of dopaminergic stimulation on the STDP time window by following a longer exponential decay in the presence of dopamine (DA^{+}) and a typical exponential decay in its absence.

In their simulations requiring the learning agent to move to a locale associated with non-stationary reward, the addition of cholinergic modulation allowed the network to rapidly unlearn the previously memorized goal locations. In contrast, the dopamine-only baseline model frequently returned to formerly learned locations of reward long after the simulation had moved their position. This is consistent with the association between acetylcholine and exploratory behaviors and the reinforcement of reward coupled with dopamine that inspired their sequential neuromodulation framework.

3.3. Weight saturation and network reconfiguration

A number of rate-based approaches to the exploration-exploitation balance have been proposed in recent neo-Hebbian RL frameworks. These efforts have largely avoided attempts at explicitly modeling additional neuromodulatory factors. One exception is that of Lew, Rey, and Zanutto (2013) which proposed a dual modulation method modeling norepinephrine rather than acetylcholine to alter the excitability of dopaminergic neurons such that exploitative strategies are only favored during periods of heightened performance (in terms of reward received). While the model produced for that work incorporated network modules and connectivity patterns with robust biological plausibility, their approach to the exploration–exploitation balance can be reduced to a form of semi-random search. This is due to the fact that during periods of lower performance norepinephrine was modeled as having an inhibitory effect on dopaminergic neurons as well as neurons in the input–output response pathway. This resulted in a heightened noise to signal ratio for response pathway neurons such that their output would fail to meet a pre-programmed threshold for activation under a winner-take-all mechanism. When the model failed to produce a response, pre-programmed responses were induced with a predetermined probability. The approach taken in Lew et al. (2013) could be improved by allowing the interactions of dopamine and norepinephrine to alter the threshold value for the winner-take-all output mechanisms during periods of exploration stimulated by norepinephrine under poor performance.

Legenstein, Chase, Schwartz, and Maass (2010) proposed an Exploratory-Hebbian (E-H) learning formulation for learning under dynamic RL tasks. Their approach combined averages of pre- and post-synaptic activations with a low-pass filter to adjust weights such that only rewards above the mean result in reinforcement of coincident activity between connected units. To stimulate exploratory behavior in the model, action selection neurons were provided input from a parameterized source of random noise drawn from a distribution with variance υ — the authors term this value as the exploration level parameter. The rate-based E-H weight update formulation is shown below.

$$\Delta w_{j,i} = \eta y_j(t)(y_i(t) - \bar{y}_i(t))(R(t) - \bar{R}(t)) \tag{26}$$

Eq. (26) replaces the direct use of post-synaptic activity in standard Hebbian plasticity with a measure on its deviation from the previous activity level mean $\bar{y}_i$, performing a simple filter on the post-synaptic excitation akin to a varying threshold separating LTP from LTD. The modulation factor corresponding to reward value is similarly filtered against its mean value $\bar{R}$. This was shown by Legenstein et al. (2010) to implicitly perform appropriate credit assignment without the use of eligibility tracing or short-term plasticity components as employed by the methods in Section 2.3.

These approximations of sliding thresholds allow for a number of mechanics for network reconfiguration at the weights. By the term "sliding", we refer to the dynamic nature of this threshold used for induction and reversal of LTP/LTD in contrast with the use of fixed threshold parameters. For example, an above average reward coincident with below average post-synaptic activity triggers LTD, while a below average reward in the same scenario results in LTP. When received rewards increase, neural units which also recently increased their activity should be given the credit for their role in attaining that heightened reward value, therefore this rule enhances the strength of their incoming synaptic weights. Those which reduced their activity prior to an increase in external reward are subject to weakening of the weighting of their input, as the input from neuron j led to reduced activity for unit i when an increase in excitation would have been appropriate. Similar arguments for the case of reduced rewards relative to past averages should be simple to assess here.

While their approach deviates little from the baseline additive noise methods used in standard neo-Hebbian RL, the ability to vary the noise level to stimulate exploratory behavior during learning is a significant factor to consider as we approach more complex methods for biasing exploration in Hebbian RL agents. The value for the variance parameter υ need not necessarily be a parameter for human specification, as it was found to function quite well over a broad range of values and only required additional weight normalization for extreme values or when employed alongside an aggressive learning rate η. An interesting variation on the E-H rule could relate the exploration level value υ to the deviation of neural activity from recent means, deviation of received rewards from recent means, or both — we explore the potential of such alterations more fully in later sections, where biological plausibility implications may inform their consideration.

Soltoggio and Stanley (2012) introduced Reconfigure-and-Saturate (RaS) Hebbian plasticity. In this work, focus was placed on the role of neuromodulation as a gating mechanism in neo-Hebbian plasticity through adaptive balancing of noise in neural activations and saturation of weight values. In their framework, allowing for the weight values to saturate (up to some maximal value) through typical neo-Hebbian reinforcement was shown to enhance the stability of the network dynamics, giving rise to purely exploitative behavioral strategies. Conversely, a combination of negative reward values and increased noise in signal transmission between connected units served to reset the learned parameters, which trended towards common values and oscillated around them for the duration of the negative signal (implicitly un-learning under negative external value). Trending towards a state of network reconfiguration, in conjunction with noise neural activity, stimulated exploration in a similarly randomized way to previously reviewed works but with the potential for selective re-learning under the exploratory regime, as demonstrated by the experiments in Soltoggio and Stanley (2012).

$$\Delta w_{j,i}(t) = w_{j,i}(t-1) + \big(C \cdot R(t) y_i(t) y_j(t-1)\big) + \xi_{j,i}(t) \tag{27}$$

Eq. (27) shows the weight update formulation corresponding to RaS Hebbian learning. As their framework incorporated learning for both excitatory and inhibitory rate-based neuron types, the variable C takes the value of $+1$ or -1 for each type. This framework additionally modeled causality through explicit propagation delays, with the learning rule relating the activity of the pre-synaptic unit at time $t-1$ with the current post-synaptic activation at time t. ξ is an additive noise to the weight calculation drawn from a uniform distribution for each synapse during updates. Through a series of experiments simulating learning under the proposed plasticity rule, network reconfiguration under LTD induced by negative rewards was found to selectively apply to connections corresponding to behavioral outcomes that required change for successful task progress. One such example discussed was re-learning to navigate due to changes in the environment's reward structure. Previously learned responses to stimuli remained intact where appropriate for attaining reward. This work was the only reviewed neo-Hebbian learning study which forewent the use of eligibility tracing or any comparable mechanism for synaptic credit assignment under delayed reward. As such, future work may consider expanding upon their methods to investigate its performance in more complex tasks and environments.

4. Motivation and modulation

Having assessed a number of recent works expanding upon the neo-Hebbian RL framework to dynamically control the balance of exploratory and exploitative behaviors in ANN and SNN agents, a number of trends in the experimental literature are apparent. Each of these proposed methods has largely focused on handling the stability–plasticity dilemma, the issue of retaining correct, stable weight solutions that remain valid in the face of alterations to task or reward structure, during phases of unlearning and re-learning. In Brzosko et al. (2017), for example, this is accomplished by flipping the sign of weight updates in a pre-determined fashion (acetylcholine levels were determined externally as a parameter) in conjunction with a reduced learning rate in the absence of dopamine reward signaling. The E–H approach proposed by Legenstein et al. (2010) performs a similar function, though in this case the flip from LTP to LTD is a dynamic mechanism based an internal history of reward signals which can be considered a simple form of intrinsic motivation (Oudeyer & Kaplan, 2009).

A core commonality to these approaches is that when the reward landscape of the task environment is either unknown or has recently changed, the models rely on phases of random rather than purposeful exploration. For more difficult RL tasks, where rewards that enable the development of exploitative strategies are sparse and potentially non-stationary, this strategy is neither computationally expedient nor derived from substantive biological evidence. In this final section, we suggest that the exploratory regime of neo-Hebbian RL should be considered a more active rather than passive process which draws inspiration from broader and more recent research on value-based learning.

4.1. Reward-prediction error hypothesis

The reward-prediction error (RPE) hypothesis posits that the dopaminergic system encodes and communicates the discrepancy between predicted and presented rewards (Bunzeck & Düzel, 2006; Pan, Schmidt, Wickens, & Hyland, 2005; Schultz, 1998). This hypothesis for the role of extrinsic reward in learning stems from experimental observations of the behavior of dopamine neurons under conditioning of reward responses to cuing stimuli. Research has shown that the dopaminergic system, when exposed to an unanticipated reward, shifts from its baseline tonic spiking mode to a phasic (highly active) mode of pulsation. This transient response shifts with continued learning of cue-reward pairings such that dopaminergic neurons become more active upon presentation of the cue than to the reward itself (Bromberg-Martin, Matsumoto, & Hikosaka, 2010). Conversely, omission of expected rewards has been associated with reduction of firing activity below the baseline tonic mode. This is considered to be analogous to a negative RPE as negative spiking activity is not supported by excitatory neurons nor is a negative firing rate possible.

The similarity between the reported dopaminergic signaling under the RPE hypothesis and the TD-error from dynamic programming has been emphasized strongly as validation for TD methods in RL (Frémaux & Gerstner, 2016; Gershman, 2018; Gläscher, Daw, Dayan, & O'Doherty, 2010; Pan et al., 2005). More recently, some experimental studies have suggested that the RPE hypothesis only captures one (albeit significant) role of dopamine neurons in regulating learning (Gardner, Schoenbaum, & Gershman, 2018; Takahashi, et al., 2017; Zhang, Lau, & Bi, 2009). Examinations of the behavior of the DA system with respect to both aversive (punishing, ''unpleasant'') stimulation as well as value-neutral (but unanticipated) events have encouraged researchers to propose a more generalized role for dopaminergic activity based on a broader view of prediction error coding in the nervous system.

Kakade and Dayan (2002) first assessed inconsistencies with the RPE hypothesis of dopamine, noting non-trivial dopaminergic responses to stimuli not associated with external rewards but with features which resemble those that the organism has previously associated with reward. Similar phasic dopamine responses were also reported for stimuli having no resemblance or direct relationship to an external reward but were novel or salient to the organism. Phasic dopamine responses can also be found, with some variation across species, in relation to specific sets of motor effects and stimuli that are irrelevant or even detrimental in goal-directed behavior. Based on these discrepancies, Kakade and Dayan (2002) proposed that the phasic response of dopamine neurons multiplexes the RPE signal with information about reward bonuses. These proposed bonuses would boost the learning process based on some internal determination of stimulus novelty as well as uncertainty in the identity and implications of unpredicted stimuli in partially observable environments. By multiplexing these proposed intrinsic bonuses, the DA system would bias the organism towards exploratory behaviors. This bias would then decrease over repeated exposure to the stimuli due to habituation unless consistently reinforced by a relation to primary reward within the environment.

Redgrave and Gurney (2006) proposed an alternative to the RPE hypothesis based on information about the sources and latency of the signals that serve as input to excite or inhibit dopaminergic neurons. The information available to the DA system in time to affect its phasic response was found to be drawn from early processing layers in sensory pathways (the authors focused on the visual system) which arrive immediately after contextual and motor efferent copy signals are received via the striatum. These early sensory processing neurons in the visual system respond prior to gaze shifts and have been shown to be sensitive to spatially localized changes in luminescence due to the sudden appearance, disappearance, or movement of salient stimuli. The latency required for the DA system to transmit a RPE-type signal is also known to be inconsistent with the required time for neural processing involved in the identification of objects as well as estimation of their reward value, which can vary greatly. This is in contrast to the phasic response of dopamine which is generally very consistent across species and shows a lower latency. The authors of Redgrave and Gurney (2006) suggested that instead of transmitting a prediction error for reward values dopamine neurons act to reinforce re-selection of actions that immediately precede an unpredicted and biologically salient event, driving the acquisition of new behaviors and sharpening the organism's ability to discern the effects of its own actions in a given environment.

Gardner et al. (2018) put forward a more general perspective on dopamine signaling, termed the sensory-prediction error (SPE) hypothesis. Under this theory, dopaminergic signals change with respect to perceptual prediction errors (defined by some theories as a neural form of surprise (Barto, Mirolli, & Baldassarre, 2013)). The corresponding reward error from the RPE hypothesis becomes a special case: biological organisms perceive rewards and punishments as stimulation and develop expectations for their timing, location, and magnitude in a similar manner to their generation of predictions about other aspects of the environment. This hypothesis is more readily reconciled with experimental results that show dopaminergic transients in response to unexpected, value-neutral stimulation as well as positive (phasic) responses to unexpected, aversive stimuli (Bromberg-Martin et al., 2010). The only discernible commonality of these responses with the stereotypical RPE phasic activity (in reaction to unpredicted rewards) is their violation of internal expectations; this more general hypothesis is further advanced by the established habituation of dopamine neurons during classical conditioning, whereby the phasic response diminishes completely as rewards become predictable.

Mirolli, Santucci, and Baldassarre (2013) were the first to reconcile the competing RPE and SPE theories of dopamine, arguing that the phasic response could be decomposed when viewing the conditioning process as involving two related learning problems: learning to effect the environment and learning how to exploit effectance. These two learning goals then divide the phasic DA response into temporary and permanent reinforcers. Learning effectance, how to act and the consequences of those actions, could be driven by the temporary phasic response generated due to surprising stimuli – a form of intrinsic reward. These temporary intrinsic rewards diminish with habituation, no longer serving to drive learning when they become predictable. The permanent portion of the response can then be understood as corresponding to the traditional RPE signal that enables the learning of exploitative behavioral responses and shifts its response to occur with stimuli that reliably predict rewards.

While it is outside the scope of this work to validate these hypotheses on the dopaminergic system, we find the arguments laid by Gardner et al. (2018) and Mirolli et al. (2013) to be particularly motivating in conjunction with similar concerns about the discrepancy between the RPE hypothesis and experimental results in both neuroscience and machine learning (Bromberg-Martin et al., 2010; Bunzeck & Düzel, 2006; Gläscher et al., 2010; Schultz, 2013; Takahashi, et al., 2017; Zhang et al., 2009). The alternative framing of dopaminergic modulation under a joined RPE-SPE hypothesis adds weight to theories that the dopamine system responds to errors generated between sensory stimulation and sophisticated predictive internal models, including the intrinsic understanding of consequence an agent must acquire to successfully act and react in a dynamic world. It additionally validates observed preferences in human and non-human animals for seeking predictable states of the environment when available, even when these states have equal expected reward value to those that are less predictable (Bromberg-Martin et al., 2010).

4.2. Free energy minimization

The free energy principle (FEP) framework established in Friston, Kilner, and Harrison (2006) pursues an overarching theory about the necessary conditions and behaviors for life, with particular developments of the theory targeting open problems in the study of mental illness, attention to salient events, and value-learning, among other areas of interest to neuroscience (Buckley, Kim, McGregor, & Seth, 2017). While much of the success of this framework as applied to these fields of study hinges on assumptions incorporated as a consequence of its derivation from probabilistic learning theories (Bayesian belief models, ensemble densities, etc.) as well as from information theoretic definitions of surprise and entropy (which are used to formally define free energy under this framework) that are not readily compatible with the established methods of neo-Hebbian RL, we find that the more abstract perspectives on learning under a free energy minimization principle may offer substantive guiding principles for developing more nuanced neo-Hebbian learning models in the future.

Under FEP theory, learning systems experience and alter their environment indirectly: some unknown generative function, taken to represent the causal environment, applies forces (light, sound, pressure, etc.) to the sensory boundaries of the organism (photoreceptor cells, etc.) that are converted into observations (stimuli that constitute the available evidence for internal models of the environment) and, conversely, the organism generates nervous responses which drive its effectors (motor cells, etc.) to exert force upon some portion of the external environment, thereby enabling it to alter its observations of and relationship to the environment. Under this theory, the fully internal portion of the organism can be considered a generative hierarchical model that actively infers causality in its relationship with the environment (Friston, Rosch, Parr, Price, & Bowman, 2017). Sensory signals are processed in increasingly abstract internal states and weighted against the system's model; sensation then becomes evidence supporting or contradicting prior expectations. This particular concept in FEP theory may be particularly compatible as an extension of the HTP framework reviewed in Section 2.3.2, which performs an RPE-only approximation of such evidence weighting.

The primary claim of the FEP framework is that adaptation, both internal with respect to belief updates and external with regard to enacting effect on the environment, requires organisms to minimize an information theoretic formulation of free energy that is, at least conceptually, highly analogous to some definitions of surprise (see Section 2.3 of Barto et al. (2013)). When confronted with unfamiliarity in the environment, organisms have two general mechanisms for reducing this surprise: (i) improve the quality of expectations generated in response to ongoing sensations such that future observations are better predicted by their internal models or (ii) act to alter the relationship they maintain with the environment to better align with prior beliefs. This latter case can result in behaviors that bring the organism to a more predictable state either by enacting a state transition (for example, retreating from the unfamiliar state to a nearby familiar one) or by acting to alter the unfamiliar state to make that state conform to the organism's belief model (such as by shivering to generate heat for homeostasis when subjected to an unanticipated drop in temperature).

Under the FEP then, learning is guided more generally by informed surprise under a perception-inference-action cycle (Friston, 2010, 2020) that influences the development of both action preferences and internal models of environmental dynamics. In line with the SPE hypothesis, preferences towards predictable outcomes in otherwise value-neutral decisions (as mentioned at the end of the previous section) emerge as surprise-minimizing behavior. By informed surprise, we refer to the updating of existing internal models due to the processing of new sensory evidence that may partially violate prior beliefs. A useful analogy for this process may be to consider what occurs when one makes an "educated guess" that is mostly correct — the true answer violates a small subset of our expectations, generating some amount of surprise which requires an update to our understanding of the problem domain, but otherwise serves to bolster the aspects of our belief model which accurately predicted most of the solution.

FEP theory, in conjunction with our earlier assessment of the reward/sensory PE hypotheses on the function of dopamine, may inform future neo-Hebbian learning frameworks in several ways. It suggests a need for a generative model of agency in RL, which has been shown elsewhere to be highly effective in gradient-based approaches (Pathak et al., 2017). It also embraces a hierarchical structuring of the brain held to be requisite for higher-order functions of perception and cognition, such as the planning of goal-directed behavior or the extrapolation of trajectories of changes in the environment, wherein downstream (deeper in terms of connective distance from the agent-environment boundary) populations of neurons adopt an increasingly associative role (operating on inputs with progressively more diverse/broad receptive fields and/or modalities) relative to their upstream counterparts.

4.3. Predictive coding

Given an ongoing stream of signals produced by the interaction of external forces upon the sensory boundary of an adaptive organism, how might the necessary information pertinent to perception be encoded? Theories of predictive coding have been suggested as a plausible solution to the problem of recovering, in terms of neural representations, the external sources of internalized sensory stimulation (Spratling, 2017).

Predictive coding theory is predicated on empirical observations about the topological properties of cortical circuitry in the brain, beginning with the layering of neural populations with distinct inter-layer (both in terms of originating and terminating) axon projections established with the Rockland–Pandya laminar rules (Rockland & Pandya, 1979). By convention, the laminar organization of the mammalian neocortex contains six layers, though this specificity holds more significance in computational approaches aiming to emulate the functionality of highly specialized neural circuitry (visual processing, for example) than to the general problem of learning to act advantageously.

The existence of these distinct connectivity patterns, both feedforward (from "lower" layers which are closer in terms of synaptic jumps towards the external boundaries of stimulation and effectance to "higher" layers which correspond to an increase in depth and abstractness of processing) as well as feedback (higher to lower), prompted Rao and Ballard (1999) to propose a basic hierarchical predictive coding scheme in the context of visual receptive fields. The intuition behind this approach is to leverage feedback predictions on the lower-level input to enforce a feedforward error coding scheme containing only useful information for learning. This selective filtering on the processing of inputs is intended to remove predictable information, corresponding to signals from the feedback circuitry, which may be considered redundant for the learning problem.

Under predictive coding theory, feedback connections convey representative predictions from higher layers in the hierarchy (which are assumed to compute more abstract functions due their increasingly broad/diverse inputs Shipp (2016)) about the expected activity of lower layers to the lower layers (typically via auxiliary neural units which aid in the calculation of discrepancy between feedback predictions and feedforward stimulation Spratling (2017)); feedforward activities then signal to higher layers the resultant mismatch between the predicted representation (from the next higher level) and the actual representation (generated in response to its inputs from the layer below it). This coding of errors by the forward flow of information in the network entails increasing sparsity of (feedforward) neural activity in proportion to the accuracy of learned representations of the environment at higher layers in the hierarchy (under the assumption that feedback connections act only to inhibit predictable forward activity, which may not reflect the full nature of biological feedback circuitry Bastos, et al. (2012)).

A hierarchical network structure integrates increasingly diverse sources of information at higher levels, leveraging prior knowledge about a given task to construct an internal topology that better enables the model to capture potential relationships between inputs/outputs of the constituent sub-networks subsumed by the hierarchy (Mavrovouniotis & Chang, 1992).

In the context of control, an agent must learn to pursue the best available course of action given only its prior experienced sequences of environmental stimuli and its internal estimate of the long-term value of actions taken therein. Conversely, that internal estimation of value is predicated on the predicted series of actions and successor states in its trajectory. Generative modeling methods with feedback in the vein of representational predictive coding theory offer an explicit means to leverage the implicit recurrence between actions and their consequences (defined here to include both environmental state features and associated external value under SPE theory). By this we mean that it may be advantageous to craft a network design with feedback that mirrors the dependency between states (both their qualities and values) and actions that is inherent to interacting with a given environment.

It will also be helpful for future research to keep in mind any temporal delays in the case of spiking neurons due to refractory dynamics as well as inherent delays in the depolarization process; the nature of these delays requires a form of predictive coding that is not only representational (in terms of one layer predicting the activities of the layer that precedes it in the hierarchy) but also temporally predictive in that feedback signaling must anticipate future feedforward spiking activity to correctly align with their representational predictive targets (Hogendoorn & Burkitt, 2019).

4.4. Research directions

Given the theoretical and experimental research discussed in the preceding sections, there are several takeaways from the literature which could inform future neo-Hebbian RL frameworks aiming to actively stimulate exploratory behaviors based on inspiration from biological learning. While the RPE hypothesis of dopamine which provides a theoretical basis for TD-learning methods is largely centered on an explicit signal to stimulate exploitation, a comparable and potentially co-existent mechanism may be required for modeling exploration as an active rather than passive process.

In non-Hebbian RL, research in this direction has largely revolved around the use of intrinsic motivators, alterations or additions to the environmental reward signal derived solely from some measure internal to the learning agent (Oudeyer & Kaplan, 2009). A particularly interesting candidate for forming such a signal may rest in the generation of sensory prediction errors. In the absence of external rewards to modulate the rate and sign of neo-Hebbian learning, the discrepancy between prediction and stimulation (assuming a generative model as discussed above) could be quantified to adjust the rate of learning during weight updates. This is in part the same effect that external reward has upon the unsupervised learning component in neo-Hebbian RL, where heightened rewards induce stronger LTP. Constructing a measure based on SPE would be simple in a neo-Hebbian framework which incorporates some form of representational predictive coding, as the feedforward flow of neural activity is assumed to only represent information which is not correctly predicted by the feedback loop.

While RL tasks are typically formulated with the intent of maximizing expected external rewards, FEP theory suggests that minimizing a SPE signal, a quantity which implicitly captures surprise, would be equivalent to evidence maximization (Friston et al., 2006). This does not require that the use of surprise as an intrinsic reward signal dampen the rate of learning — logically, heightened surprise should occur when the model has more available information to learn about its environment and the consequences of its actions therein.

Future research should investigate further the biological basis for the construction of such a measure as well as the potential dynamics for its use as an intrinsic motivator alongside the standard external reward signal. We hypothesize that such a mechanism for extending neo-Hebbian RL to motivate non-random exploratory behaviors during learning can be formulated by incorporating one or more additional terms in the learning rule. This expansion on the neo-Hebbian RL framework will require the construction of a potentially complex neural network module, in the vein of representational predictive coding theory, to generate values internally for such additional modulatory factors.

5. Conclusions

Neo-Hebbian learning models which dynamically adjust their exploration–exploitation balance have demonstrated improved exploratory performance in the context of RL, particularly where environments of interest contain sparse or non-stationary extrinsic reward structures. While these improvements are non-trivial, their methods remain relatively simple in comparison to recent advances in the domain of deep RL. We find that the dominant approach employed by the reviewed neo-Hebbian methods, gating and scaling effects applied to Hebbian update rules via perturbations of the learning rate parameter, lacks a clear path for substantive improvement without re-conceptualization. Alternative mechanisms for modulating this dynamic, whether drawn from research in neuroscience or advances in deep RL, have the potential to open promising new lines of work for this class of learning model.

We find that there is ample inspiration for future work in this direction to be found in recent research in neuroscience, particularly works relating dopamine as a neuromodulator influenced by sensory prediction errors generated in response to active sampling of an organism's environment. We argue that this extension to the reward-prediction error hypothesis, which inspired TD learning methods, allows for a more realistic modeling of exploratory behaviors. The broader research in this context suggests that representational predictive coding, when applied in such a way to align with the surprise minimization principles of the FEP framework, offers a promising direction for future work in expanding upon the neo-Hebbian RL model to more plausibly replicate the exploration–exploitation dynamics observed in animals. We expect that further research in biological and computational neuroscience will advance our understanding of the motivational factors in value-driven learning and have significant impact to the design of next-generation exploration–exploitation dynamics for artificial learning systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Apicella, A., Donnarumma, F., Isgrò, F., & Prevete, R. (2021). A survey on modern trainable activation functions. *Neural Networks*, *138*, 14–32.

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., & Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Frontiers in Psychology*, *5*, 985.

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

Barto, A. G., Singh, S., Chentanez, N., et al. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning* (pp. 112–119). Piscataway, NJ.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834.

Brzosko, Z., Zannone, S., Schultz, W., Clopath, C., & Paulsen, O. (2017). Sequential neuromodulation of Hebbian plasticity offers mechanism for effective reward-based navigation. *ELife*, *6*.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, *51*(3), 369–379.

Chentanez, N., Barto, A., & Singh, S. (2004). Intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*, *17*.

Dong, Z., Gong, B., Li, H., Bai, Y., Wu, X., Huang, Y., et al. (2012). Mechanisms of hippocampal long-term depression are required for memory enhancement by novelty exploration. *Journal of Neuroscience*, *32*(35), 11980–11990.

Feldman, D. E. (2012). The spike-timing dependence of plasticity. *Neuron*, *75*(4), 556–571.

Fink, C. G., Murphy, G. G., Zochowski, M., & Booth, V. (2013). A dynamical role for acetylcholine in synaptic renormalization. *PLoS Computational Biology*, *9*(3).

Fourcaud-Trocmé, N., Hansel, D., van Vreeswijk, C., & Brunel, N. (2003). How spike generation mechanisms determine the neuronal response to fluctuating inputs. *Journal of Neuroscience*, *23*(37), 11628–11640.

Frémaux, N., & Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in Neural Circuits*, *9*.

Frémaux, N., Sprekeler, H., & Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Computational Biology*, *9*(4).

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2).

Friston, K. (2020). Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, *96*.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal de Physiologie (Paris)*, *100*(1), 70–87.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, *77*, 388–402.

Gardner, B., & Grüning, A. (2013). Learning temporally precise spiking patterns through reward modulated spike-timing-dependent plasticity. In *International Conference on Artificial Neural Networks* (pp. 256–263). Springer.

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891).

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *Journal of Neuroscience*, *38*(33), 7193–7200.

Gerstner, W. (1990). Associative memory in a network of 'biological' neurons. *Advances in Neural Information Processing Systems*, *3*.

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: Experimental support of NeoHebbian three-factor learning rules. *Frontiers in Neural Circuits*, *12*, 53.

Gerstner, W., Ritz, R., & Van Hemmen, J. L. (1993). Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological Cybernetics*, *69*(5), 503–515.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4).

Golden, R., Rossa, M. A., & Olayinka, T. J. (2016). Parametrization of neuromodulation in reinforcement learning.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gordon, G., Dorfman, N., & Ahissar, E. (2013). Reinforcement active learning in the vibrissae system: Optimal object localization. *Journal de Physiologie (Paris)*, *107*(1–2), 107–115.

Gregor, M., & Spalek, J. (2014). Novelty detector for reinforcement learning based on forecasting. In *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics* (pp. 73–78).

Hansel, D., & Mato, G. (2001). Existence and stability of persistent states in large neuronal networks. *Physical Review Letters*, *86*(18), 4175.

Hao, Y., Huang, X., Dong, M., & Xu, B. (2020). A biologically plausible supervised learning method for spiking neural networks using the symmetric STDP rule. *Neural Networks*, *121*, 387–395.

Harlow, H. F., Harlow, M. K., & Meyer, D. R. (1950). Learning motivated by a manipulation drive. *Journal of Experimental Psychology*, *40*(2), 228.

He, H., & Zhong, X. (2018). Learning without external reward. *IEEE Computational Intelligence Magazine*, *13*(3), 48–54.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hoerzer, G. M., Legenstein, R., & Maass, W. (2012). Emergence of complex computational structures from chaotic neural networks through reward-modulated Hebbian learning. *Cerebral Cortex*, *24*(3), 677–690.

Hogendoorn, H., & Burkitt, A. N. (2019). Predictive coding with neural transmission delays: A real-time temporal alignment hypothesis. *ENeuro*, *6*(2).

Hu, H., Song, S., & Huang, G. (2019). Self-attention-based temporary curiosity in reinforcement learning exploration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1–12.

Huang, X., Wu, W., Yin, P., & Qiao, H. (2017). Improving learning efficiency of recurrent neural network through adjusting weights of all layers in a biologically-inspired framework. In *2017 International Joint Conference on Neural Networks* (pp. 873–879).

Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. Appleton-Century.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *BMC Neuroscience*, *8*.

Jawed, S., Grabocka, J., & Schmidt-Thieme, L. (2020). Self-supervised learning for semi-supervised time series classification. *Advances in Knowledge Discovery and Data Mining*, *12084*, 499.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4), 549–559.

Kosko, B. (1986). Differential Hebbian learning. In *AIP conference proceedings (vol. 151)* (pp. 277–282). American Institute of Physics.

Kuriscak, E., Marsalek, P., Stroffek, J., & Toth, P. G. (2015). Biological context of hebb learning in artificial neural networks, a review. *Neurocomputing*, *152*, 27–35.

Kuśmierz, Ł., Isomura, T., & Toyoizumi, T. (2017). Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology*, *46*, 170–177.

Lapique, L. (1907). Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal of Physiology and Pathology*, *9*, 620–635.

Latham, P. E., Richmond, B., Nelson, P., & Nirenberg, S. (2000). Intrinsic dynamics in neuronal networks. I. Theory. *Journal of Neurophysiology*, *83*(2), 808–827.

Lee, D.-H., Zhang, S., Fischer, A., & Bengio, Y. (2015). Difference target propagation. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 498–515). Springer.

Legenstein, R., Chase, S. M., Schwartz, A. B., & Maass, W. (2010). A reward-modulated Hebbian learning rule can explain experimentally observed network reorganization in a brain control task. *Journal of Neuroscience*, *30*(25), 8400–8410.

Lew, S., Rey, H. G., & Zanutto, B. S. (2013). Neuronal mechanisms underlying exploration-exploitation strategies in operant learning. In *The 2013 international joint conference on neural networks* (pp. 1–6).

Machado, M. C., Bellemare, M. G., & Bowling, M. (2020). Count-based exploration with the successor representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5125–5133.

Malenka, R., & Bear, M. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*, 5–21.

Markram, H., Gerstner, W., & Sjöström, P. J. (2011). A history of spike-timing-dependent plasticity. *Frontiers in Synaptic Neuroscience*, *3*.

Mavrovouniotis, M. L., & Chang, S. (1992). Hierarchical neural networks. *Computers & Chemical Engineering*, *16*(4), 347–369.

Mirolli, M., Santucci, V. G., & Baldassarre, G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: A simulated robotic study. *Neural Networks*, *39*, 40–51.

Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A., Thorpe, S. J., & Masquelier, T. (2018). Combining STDP and reward-modulated STDP in deep convolutional spiking neural networks for digit recognition. *1*, arXiv preprint arXiv:1804.00227.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Oudeyer, P.-Y., & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*.

Ozturk, I., & Halliday, D. M. (2016). Mapping spatio-temporally encoded patterns by reward-modulated STDP in spiking neurons. In *IEEE symposium series on computational intelligence*.

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*(26), 6235–6242.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th international conference on machine learning (vol. 70)* (pp. 2778–2787).

Paugam-Moisy, H., & Bohte, S. (2012). Computing with spiking neuron networks. In *Handbook of natural computing* (pp. 335–376). Springer Berlin Heidelberg.

Pogodin, R., Corneil, D., Seeholzer, A., Heng, J., & Gerstner, W. (2019). Working memory facilitates reward-modulated Hebbian learning in recurrent neural networks. arXiv preprint arXiv:1910.10559.

Porr, B., & Wörgötter, F. (2003). Isotropic sequence order learning. *Neural Computation*, *15*(4), 831–864.

Potjans, W., Diesmann, M., & Morrison, A. (2011). An imperfect dopaminergic error signal can drive temporal-difference learning. *PLoS Computational Biology*, *7*(5).

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975.

Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*(1), 3–20.

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, *19*(3), 166.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Schembri, M., Mirolli, M., & Baldassarre, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *2007 IEEE 6th international conference on development and learning* (pp. 282–287).

Schmidhuber, J. (1991). Curious model-building control systems. In *[Proceedings] 1991 IEEE international joint conference on neural networks (vol. 2)* (pp. 1458–1463).

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Schultz, W. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology*, *23*(2), 229–238.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, *40*(6).

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, *7*.

Shouval, H., Wang, S., & Wittenberg, G. (2010). Spike timing dependent plasticity: A consequence of more fundamental learning rules. *Frontiers in Computational Neuroscience*, *4*.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, *7*, 53040–53065.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82.

Soltoggio, A. (2015). Short-term plasticity as cause–effect hypothesis testing in distal reward learning. *Biological Cybernetics*, *109*(1), 75–94.

Soltoggio, A., & Stanley, K. O. (2012). From modulated Hebbian plasticity to simple behavior learning through noise and weight saturation. *Neural Networks*, *34*, 28–41.

Soltoggio, A., & Steil, J. J. (2013). Solving the distal reward problem with rare correlations. *Neural Computation*, *25*(4), 940–978.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction* (1st ed.). MIT Press.

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement learning: an introduction*. The MIT Press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, *95*(6), 1395–1405.

Tetzlaff, C., Kolodziejski, C., Markelic, I., & Wörgötter, F. (2012). Time scales of memory, learning, and plasticity. *Biological Cybernetics*, *106*, 715–726.

Tuckwell, H. C. (1988). *Introduction to theoretical neurobiology: linear cable theory and dendritic structure (vol. 1)*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in neural information processing systems (vol. 30)*. Curran Associates, Inc..

Wang, J., Belatreche, A., Maguire, L., & McGinnity, T. M. (2014). An online supervised learning method for spiking neural networks with adaptive structure. *Neurocomputing*, *144*, 526–536.

Yusoffa, N., & Grüning, A. (2012). Biologically inspired temporal sequence learning. *Procedia Engineering*, *41*, 319–325.

Zannone, S., Brzosko, Z., Paulsen, O., & Clopath, C. (2018). Acetylcholine-modulated plasticity in reward-driven navigation: a computational study. *Scientific Reports*, *8*(9486).

Zappacosta, S., Mannella, F., Mirolli, M., & Baldassarre, G. (2018). General differential Hebbian learning: Capturing temporal relations between events in neural networks and the brain. *PLoS Computational Biology*, *14*(8).

Zenke, F., & Ganguli, S. (2018). Superspike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, *30*(6), 1514–1541.

Zhang, J.-C., Lau, P.-M., & Bi, G.-Q. (2009). Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. *Proceedings of the National Academy of Sciences*, *106*(31), 13028–13033.