UNIVERSITEIT VAN AMSTERDAM

Learning from the brain

Exploring neoHebbian plasticity and memory replay in reinforcement learning

by

Paolo Umberto Agliati

November 2022

A Dissertation submitted in part fulfilment of the
Research Master:
Brain and Cognitive Sciences
Universiteit van Amsterdam

Supervised by:
Renato Farinha Duarte,
Donders institute for brain, cognition and behaviour
and
Herke van Hoof,
AMLab, Universiteit van Amsterdam

Contents

Abstract

Learning in realistic environments remains a novel domain within the fields of artificial intelligence and neuroscience from which researchers could step towards unveiling the principles of brain functionality and advancing models in their capabilities. Focusing on reinforcement learning, this review addresses the use of currently known principles from computational neuroscience and biology to tackle the main challenges in the field, with particular emphasis on improving inductive biases and preventing catastrophic interference. The current work proposes the interaction of two bio-inspired features, namely neuromodulation of synaptic plasticity and memory replay as a promising avenue to face these obstacles and to provide artificial systems with a human-like learning process. The majority of studies support the idea that neuromodulation and memory replay are highly intertwined events, both needed for the learning process in humans. Moreover, attempts at including these features in reinforcement learning agents proved to be beneficial for the model's performance in ecological settings. In this light, we explore how spike timing can represent a valid substrate on which to implement both bio-plausible characteristics. The framework of spiking neural networks is often utilised to represent and remember relevant features of specific environments, especially in dynamic contexts with sparse rewards. However, the literature offers different approaches to model both neuromodulation and memory representation, with varying degrees of cross-compatibility. Finally, we explore the challenges in implementing these bio-inspired systems, allowing computational models to aid current research, including neuromorphic hardware development, robotics, and neuroscience.

1 Introduction

Characterizing the mechanisms behind the phenomenon of learning especially in complex, realistic environments, has always been a central question in neuroscience. Behavioral studies of decision-making in humans and animals were the first contributions to the understanding of such a fundamental yet complex process. Investigations on the subject started through the theories of Pavlovian (classical) conditioning (Pavlov, 1928) and instrumental (operant) conditioning (Thorndike, 1927; Rodhom and Tolman, 1950). Empirically, these initial studies provided standardized means to quantify the processes involved in behavioral decision-making and the role of reward and punishment. Theoretically, aided by the nascent field of mathematical psychology and by significant advancements in control engineering (e.g., dynamic programming, Bellman, 1952), these studies provided important foundations for modern reinforcement learning (RL). For example, early mathematical formulations of classical and instrumental conditioning (the Rescorla-Wagner rule, Rescorla and Wagner, 1972) established the basis of what is now known as temporal difference (TD) learning. The ability to quantify, mathematically formalize and validate behavioral (and subsequently neural) data has since become central to our understanding of animal behavior, learning and cognitive computing. RL is the major modeling construct currently available in this domain, and it is often used to model the behavior of an agent in a given learning environment (Dayan and Niv, 2008). In these models, agents learn the utility or value of certain choices based on different inputs they receive as a consequence of those choices (mainly reward inputs). In the study of reward-driven learning (O'Doherty et al., 2015; Sutton and Barto, 2018), RL paradigms are employed to describe how the outcomes of certain actions drive the subjective values the learning agent holds for these actions, as is the case in Q-learning (Watkins, 1989; Watkins and Dayan, 1992), one of the most common frameworks used in the field. This conceptualization of learning differs from the other two central theories of learning: supervised and unsupervised learning. In supervised learning (SL), the agent (or algorithm) is presented with the appropriate response (e.g. the correct output in a classification task) during the training phase; while in unsupervised learning (UL) only the inputs are presented, and the agent learns to build a progressively better data-driven representation of the structure of such input without any additional information (as is the case for clustering algorithms, for example). On the one hand, the paradigm of RL shares similarities with both: the presence of a guidance is in common with SL (albeit in RL, the desired outcome is not shown to the agent, but a measure of distance from the target drives the agent towards a goal), while the ability of the agent to self-organize useful representations is shared with UL. On the other hand, RL differs fundamentally from these two, since the agent learns by discovering actions that increase reward, given the setting in which it operates. This means the agent acquires some form of representation of the environment (and/or its reward structure), but just through the active exploration (and/or exploitation) of the possible actions and states available in that specific environment. Thus, central questions in RL as a theory of animal learning revolve around

the way in which such representations are formed, how much they can be generalized and how should they inform an action selection policy. These same questions can become especially meaningful for this field when we consider realistic environments, which may be partially observable, noisy, uncertain and dynamic. Additionally, representations of reward structure and/or the dynamics of environmental state transitions constitute an internal model of the task (and of its setting) which the agent must acquire by experience and whose internal structure may become extremely complex. RL algorithms can in fact be divided into model-free (MF) and model-based (MB). These two classes of algorithms are distinguished by the extent (and the modality) with which they allow the agent to form a representation of the environment (Collins and Cockburn, 2020). MF algorithms do not maintain an explicit model of the environment they interact with. Instead, they allow the agent to have a record of the acquired rewards in such environment. This means a MF agent learns a functional mapping between reward inputs and the corresponding behaviors in its output. On the opposite end, MB algorithms allow the agent to form explicit representations of the environment, which may include the state and action spaces as well as the reward functions and the transition functions (Collins and Cockburn, 2020; Moerland et al., 2023). Both approaches have their own shortcomings. MB RL can come up with more elaborate and flexible action plans but doing so often proves to be computationally expensive or intractable, especially when the environment is complex and dynamic (Moerland et al., 2023). MF algorithms are more tractable, but they do not adapt as easily to changes in the environment or to new settings (Calisir and Pehlivanoglu, 2019). Moreover, some problems affect both approaches and are generally present in the RL paradigm. One example of a prevalent issue is catastrophic interference, which occurs when a learning system loses previously acquired associations after the integration of new knowledge from the environment. Another important issue occurs when networks have a weak inductive bias: having weak assumptions about any given environment allows the agent to adapt to a bigger number of challenges but requires much more data during the training process (Botvinick et al., 2019). Beyond the strictly modeling-related issues, MB and MF are still active areas of debate when it comes to the study of learning in the human brain. The brain is thought to be a hybrid learning system, able to exploit both MB and MF learning paradigms. These two are integrated by relying on goal-directed actions (result of MF learning) early on in the pres-

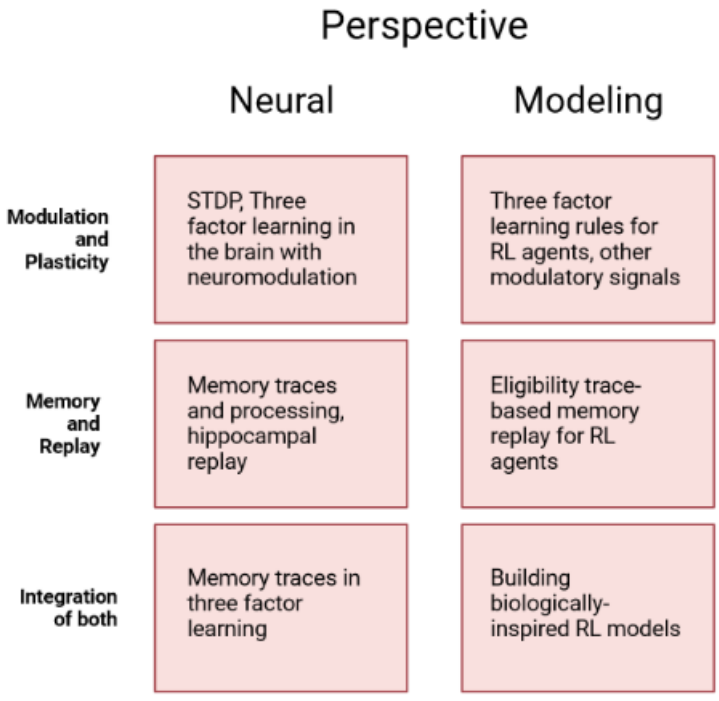

Figure 1: Scheme of the content and structure of this review (RL: reinforcement learning; STDP: spike-timing-dependent plasticity). The neural perspective will be discussed in section 2.1, while the efforts to model it will be covered in section 2.2.

ence of novel stimuli and settings, while favoring habitual actions (result of MB learning) after more extensive training (Daw et al., 2005; O'Doherty et al., 2015). This optimizes the way in which representations are formed during the learning process, since a more detailed representation will be acquired only if the environment requires it from the learning agent (Wilson and Niv, 2012). Building such simple, generalized representations constraints the complexity of the problem state-space and speeds up learning by increasing computational efficiency, since the brain avoids having to re-learn recurring environmental patterns (O'Doherty et al., 2015). The exact details of how MB and MF learning are combined in the human brain are still being studied and understood. For a more thorough account of the currently known biological and anatomical correlates of these, see section 2.1.1. A deeper look into this dichotomy and in how we can move past it remain central themes both for the improvement of the current model's performance, and for increasing their biological plausibility, advancing computational neuroscience in the ability to explain learning in realistic, ecological settings. To illustrate, MF algorithms become particularly important in realistic environments where an explicit internal representation or a complete model is unattainable. Beyond the MB/MF dichotomy, however, there are other important factors an agent needs to consider when operating in ecological settings: observability, volatility, and sparse, dynamic reward structure. To address some of these complications and solve RL problems in high-dimensional environments with complex reward structures, deep artificial neural networks (ANNs) have been successfully employed to learn approximate value functions (Arulkumaran et al., 2017). This approach, collectively known as Deep-RL has fundamentally revolutionized the field, allowing RL algorithms to tackle previously unsolvable problems. From a neuro-cognitive standpoint, this combination is also conceptually insightful as it may hint on potential solutions employed by the brain to solve complex cognitive problems. In particular, the hierarchical nature of internal representations in deep ANNs combined with the ability to approximate arbitrary value functions suggests similar processes may be at play in the mammalian brain (Botvinick et al., 2020). Additionally, the shortcomings of this approach may find useful solutions in the brain's mechanisms. Let us consider, when it comes to generalization and adaptation to different types of environments, weak inductive bias, which seems to be the main challenge for deep-RL. This characteristic is incompatible with the sparsity of inputs that distinguish realistic from controlled settings. Once again, the brain's solution to weak inductive biases is an active area of research. The mammalian brain is an evolved architecture and the process of evolution by natural selection is itself a learning process (postulated to be "learning by direct-fit" by Hasson et al., 2020). For instance, the iterative optimization process across generations of evolving species finds a parallel in the iteration over samples we find in networks of neurons. In both cases, the gradual adjustment of parameters improves the fit of the organism. Natural organisms in general and brains in particular are overparameterized systems, where multiple tunable parameters are potential learning targets. However, the type and nature of their modifications is subject to important physical and biological constraints. For example, cortical circuits exhibit detailed balance between excitation and inhibition (the E/I balance), which assures that the average weight strength is scaled according to the number and type of incoming synapses (van Vreeswijk and Sompolinsky, 1997), i.e. if excitation becomes too strong, inhibition needs to be re-balanced. This structural constraint allows any given neuron to elicit spikes, since it modulates its input based on the number of incoming synapses and illustrates the fact that biological learning mechanisms are not allowed to freely tune system parameters, but are constrained by the system's architecture and design. This overlap between the structural elements of a network and its functions are at the base of the brain's ability to optimize inductive biases, since it allows to efficiently store learned rules and assumptions about the current environment. Candidate solutions to strengthen inductive biases in RL agents present varying degrees of inspiration from the study of the brain. Some of them tackle this problem (as well as other key issues in the RL framework) by aiming to move past the MB-MF dichotomy without taking any direct inspiration from biology. This often entails allowing the agent to form some sort of representation of the environment without having an explicit, model-based representation. A good illustration of this approach is given by Successor Representation algorithms (SR), in which the value function is stored along with a predictive but not explicit representation. This makes SR a hybrid between the computational efficiency of MF and the flexibility of MB RL (Momennejad et al., 2017; Russek et al., 2017). Other proposed solutions aim at getting closer to the brain's functioning without directly modeling any of its features. Examples include a slow meta-learning algorithm that aids generalization on top of a standard RL algorithm; for more detailed information on this approach see Botvinick et al., 2019. Such feature resembles the role of prefrontal cortex (PFC) in learning, which aids the activity of the dopaminergic and cortico-striatal systems (O'Doherty et al., 2015). The hierarchical architecture of the cor-

tex also lends itself well to the implementation of a Bayesian model, which could aid in carrying out "imprecise computations", useful to face sparse reward environments (Findling et al., 2020). Lastly, some methods to overcome the issues with modern RL include taking more direct inspiration from the process of learning that occurs in the brain by implementing biologically-plausible features and explicitly modeling the mechanisms. Arguably, this is the most explanatory approach as model parameters and variables can be directly mapped onto known biophysical processes. In this review, we discuss two main approaches to construct biologically-plausible RL models, and whether they can be exploited in synchrony to maximize the efficiency and the performance of RL agents. The first approach aims at building plasticity in the network to have a better method of updating what is learned, in a contingent way with what is being rewarded to the agent. A second, biologically inspired feature consists in implementing internal generative replay to aid generalization in changing environments.

2 Literature Overview

2.1 Learning in the brain

2.1.1 Learning modalities in animals and humans

In order to make sense of which biological features we want to base our models on, we will now analyze the components and modalities of the learning processes in the mammalian brain, and the computational bases that could be of interest for the improvement of such models. Differently from most artificial systems, the brain is thought to implement all the aforementioned paradigms of learning (SL, UL, RL) in different regions and sub-networks, and rarely employs one single modality in isolation. Supervised learning in the human brain is thought to be implemented in a network of regions including the prefrontal cortex (involved in error-correction during rule-based learning) and the parietal cortex (implicated in updating value representations during decision-making). Depending on the task at hand, additional structures like the hippocampus or the cerebellum can also be involved. One example of the usage of SL in the human brain is the acquisition of new motor skills, such as learning to play a musical instrument. As a person learns to play an instrument, the brain receives feedback on the accuracy of movements and adjust them in accordance with the desired output. In this process, the cerebellum gives the direct supervisory signal, constituting a form of SL, and communicates with the primary motor cortex and the premotor cortex in the planning, execution, and formation of internal models of movement (Krakauer and Shadmehr, 2006; Doyon et al., 2002). An instructing signal able to constrain the behavior of the learning agent does not need to originate from a supervised learning process. An illustration of this concept in the brain comes from evolution of innate behaviors. Animals show a diverse set of innate behaviors able to contribute to the organism's fitness in its surroundings. This means that learning from experience (in the span of an individual's lifetime) is just one of the brain's methods to aid adaptation, but if we refer to learning as "encoding statistical regularities from the outside world" (Zador, 2019) we cannot ignore the strict rules that emerge from the learning process of evolution. In his work, Zador, 2019 argues that although evolution acts on the brain wiring in an indirect fashion, its effects on the genome and, consequently, on brain structure and behavior, are a strong example of a learning signal coming from a non-supervised process.

As previously stated, in unsupervised learning a model is trained to discover patterns and relationships in a dataset without the use of labeled examples or explicit supervision (Bengio et al., 2013). UL has also been found to play a significant role in the human brain for the development of sensory systems, as is the case for the infants' ventral visual stream (Higgins et al., 2016; Bremner et al., 2015), and for the formation of internal representations of the stimuli conveyed by such systems (Kohonen, 2001; Zador, 2019). During the first few months of life, the human brain finds itself in an environment with a vast amount of previously unexperienced information; to make sense of this information, the brain must learn to extract meaningful patterns and relationships from its input and does so by grouping related stimuli together and forming categories, or clusters (Kohonen, 2001). Similarly, when we encounter a new experience, the brain must select and store what is relevant about it in a way that allows it to establish a coherent sensory representation of its surroundings. Although some responses to visual stimuli are innate and accomplished via genetically determined behavioral rules, the progressive construction of an internal sensory representation in a developing animal is thought to be carried out via UL (Zador, 2019). There is also evidence that UL mechanisms may be involved in higher cognitive functions, such as language processing (Bengio et al., 2013). In this field, research has shown that the brain is capable of learning the structure of a new language without explicit instruction (Li and Zhao, 2013) suggesting that UL mechanisms may be involved in this process.

Lastly, anatomical and functional correlates of RL can be found in the human brain. An example is

the striatum, a brain region involved in motor control and reward processing, which is activated during RL tasks (Averbeck and O'Doherty, 2022). Instances of other areas associated with RL in the brain include the amygdala, which is involved in processing emotional information and learning from punishment (performing cost and benefit integration), and the PFC, which is involved in working memory and executive function in decision-making (Dixon and Dweck, 2022). Specifically, the orbitofrontal cortex appears to be the most active prefrontal area in learning from feedback (Groman et al., 2019; Costa and Averbeck, 2020). Interestingly, MF and MB RL also find their correlates in areas and substructures of the brain. The dorsal striatum (DS) is activated during tasks that involve the use of a learned value function (Jessup and O'Doherty, 2011). Thanks to the wide span of regions it communicates with (including the hippocampus, for spatial navigation, the amygdala, for emotional information, but also the motor cortex, which is responsible for initiating movement), the DS (in particular the dorsolateral striatum) is believed to contribute to the goal of maximizing future rewards in a MF fashion (van der Meer et al., 2010; Skelin et al., 2014; Geerts et al., 2020). Furthermore, several studies investigated the role of the ventral striatum (VS) and the PFC in MB RL (McDannald et al., 2011; Daw et al., 2011). The VS is thought to contribute to the learning of action-outcome associations: substructures of this region seem to cover different roles in updating the model of the environment based on new information (for instance probabilistic versus immediate rewards, or feedback-independent information, Filimon et al., 2020). Additionally, the action of neuromodulators like dopamine seems to influence the balance between model-based and model-free approaches by acting on the VS and the lateral PFC to exert a form of behavioral control (Deserno et al., 2015). The upcoming section will discuss in more details the computational relevance of neuromodulation in the brain, and why such event is crucial to understand information processing in RL.

2.1.2 RL, plasticity, neuromodulation

Extensive research on learning and synaptic plasticity stemmed from the first studies by Donald Hebb, according to which neurons in close proximity that are frequently co-active will form stronger synapses (Hebb, 1950). *Hebbian learning* principles find strong experimental support in their proposed physiological correlates: long-term potentiation (LTP) and long-term depression (LTD) (Bliss and Lømo, 1973). These two events modulate the strength of synapses according to the correlation and timing of activations of pre- and post-synaptic neurons (synchronous firing causes LTP while asynchronous firing causes LTD). Different spiking behavior over time seems to adjust synaptic strength by causing an increase or a decrease of exposed receptors for the neurotransmitter involved in that synapse. More details about the mechanisms of post-synaptic receptor exposure in LTP and LTD can be found in the works of Zamanillo et al., 1999; Giese et al., 1998; and Nabavi et al., 2014. This means learning as a phenomenon must be linked to permanent (yet plastic) changes in synaptic structure over time, but the plausible mechanism for the brain to implement such changes and make them contingent on what is being learned is still an active area of research (Pawlak, 2010; Anisimova et al., 2022). Spike-timing-dependent plasticity (STDP), as a particularly prominent biological implementation of Hebbian learning, could represent a learning rule able to support the phenomena of LTP and LTD in brain circuitry. However, spike-timing alone cannot accomodate all empyrical observations on LTP/LTD and many additional factors such as dendritic location, local interactions between neighbouring synapses, and local signaling molecules could influence plasticity as well (Brzosko et al., 2019). Furthermore, the role of STDP as a learning rule requires further research, since many of the findings do report the presence of STDP *in vitro*, under specific stimulation protocols that have been argued to be implausible. Additionally, *in vivo* circuits have significant complications in the rules for synaptic plasticity (Caporale and Dan, 2008). Crucially, in this conceptualization of plasticity, spike timing is a primary factor for learning to occur, since it drives the strengthening and weakening of connections involved in forming a representation of the environment and of its rewards. However, when considering the macro-circuitry that most directly approximates the implementation of reinforcement learning in the brain: the cortico-striatal system (Fisher et al., 2017; Badre and Frank, 2012; Frank and Badre, 2012), it is evident how at this scale, spike timing represents a necessity for the reinforcement signal, but is not sufficient to describe how learning occurs. A purely Hebbian perspective does not seem to be sufficient, and additional learning rules need to be taken into consideration. Specifically, the timing of these spiking events (milliseconds) as well as the dependence of learning purely on pre- and post-synaptic correlations, seems to be incompatible with the time scale with which behavior occurs (usually from hundreds of milliseconds to few seconds) and with the need for an external reward to be integrated into the learning process. An additional element is proposed to bridge this gap in biological systems: neuromodulation (Br-

zosko et al., 2019). Plasticity in larger brain areas is usually under the control of neuromodulatory factors, which act at a slower time scale and often influence the connectivity of whole networks instead of affecting specific synapses (Zoli et al., 1998; Matsuda et al., 2009). In the brain, neuromodulators are a vast family of signaling molecules which includes the classes of catecholamines, as well as serotonergic and histaminergic factors. These usually act at a distance and are produced by cell bodies placed in separate nuclei, with long projections that reach the areas of interest (for instance, the thalamus or the cortex). A common example of neuromodulator is represented by the dopamine molecule and its influence on cortical networks to signal reward prediction errors (RPEs; Björklund and Dunnett, 2007). The introduction of neuromodulators in our conceptualization of learning is not only an attempt to blindly mimic biological systems, but it carries great significance in developing a biologically plausible learning rule: where pre- and post-synaptic spiking activity were the first two factors to determine the state of synapses in a network, neuromodulation will be referred to as the third factor, which has the unique and crucial role of carrying information about the success of an action (and the associated reward), shaping the representation of a certain environment and bridging the relevant timescales. Differently from a purely Hebbian account on learning, this updated learning rule would assume the following form:

$$\Delta W_{i,j} \propto F(pre, post, \mathrm{MOD}) \tag{1}$$

where $W_{i,j}$ represents the weight of a synapse in the network, the change of which is a function (F) of the spiking frequencies or timings of pre- and post-synaptic neurons (the local parameters "pre" and "post") and of the concentration or presence of neuromodulation (the global parameter "MOD"). Taking into account a three-factor learning rule becomes extremely important when considering reinforcement learning, since the information relative to the reward shapes the internal representation and ultimately, the behavior of the agent. This carried information will then be used by the brain to adjust synaptic weights in the network according to the rules determined by the first two factors. Therefore, a purely Hebbian account of pre- and post-synaptic spike timing does not directly cause synaptic plasticity, but rather makes some synapses eligible to be reinforced by the concurrent action (the events can be separated by a window of approximately 1 second) of neuromodulation (Fisher et al., 2017). This concept, known as eligibility traces, will be further explored in section 2.1.3. In their review, Triche et al., 2022 explore the present efforts in integrating reward-contingent STDP (R-STDP, here referred to as neoHebbian Learning) in a RL agent, and how to include neuromodulation in the same model to obtain a more efficient and biologically plausible method of learning. More about the contribution of biologically inspired learning rules for plasticity and neuromodulation in the modeling field can be found in section 2.2.2.

2.1.3 Learning and memory replay

In broad strokes, memory in the brain can be described as an adaptive phenomenon linked to the emergence of neural activity able to last longer than the input which caused it (Chaudhuri and Fiete, 2016). The mechanisms of memory consolidation are likely ubiquitous in the brain and shared among every self recurrent network. They involve a vast network of intracellular pathways which convert electrical activity to the states of different chemical variables. One of the (computationally most explored) consequences is known as persistent activity, resulting in a population of neurons being able to sustain elevated firing rates in the absence of external input. When an input is given to the network, the activity over time of its neurons increases, and the self recurrent projections sustain this increase creating a positive feedback, as illustrated in Figure 2. From the perspective of neural dynamics, persistent activity is a distinguishing feature of a system able to reach a bistable state. In this case, the network will have access to two different regimes of firing rate over time (attractors of the network's dynamical system, Figure 2), depending on whether the information in input is being retained or not. Of course, spike timing and network plasticity are also conventionally linked to the ability of the network to process information, which means that in the brain, processing information and storing it are two operations carried out by the same computing structure, a feature that greatly aids the efficiency of biological neural networks. A deeper dive into the computational principles of memory in biological neural networks can be found in the works of Chaudhuri and Fiete, 2016 and Gallistel and King, 2009.

In addition to this feature, humans as learning agents have another important element dedicated to information consolidation: the hippocampus. This specialized region is implicated in the consolidation of semantic and episodic memory (Burgess et al., 2002; Duff et al., 2020). Taking spatial navigation as an example, the hippocampus is equipped with functionally specialized cell types such as place cells, grid cells and head direction cells, which are acti-

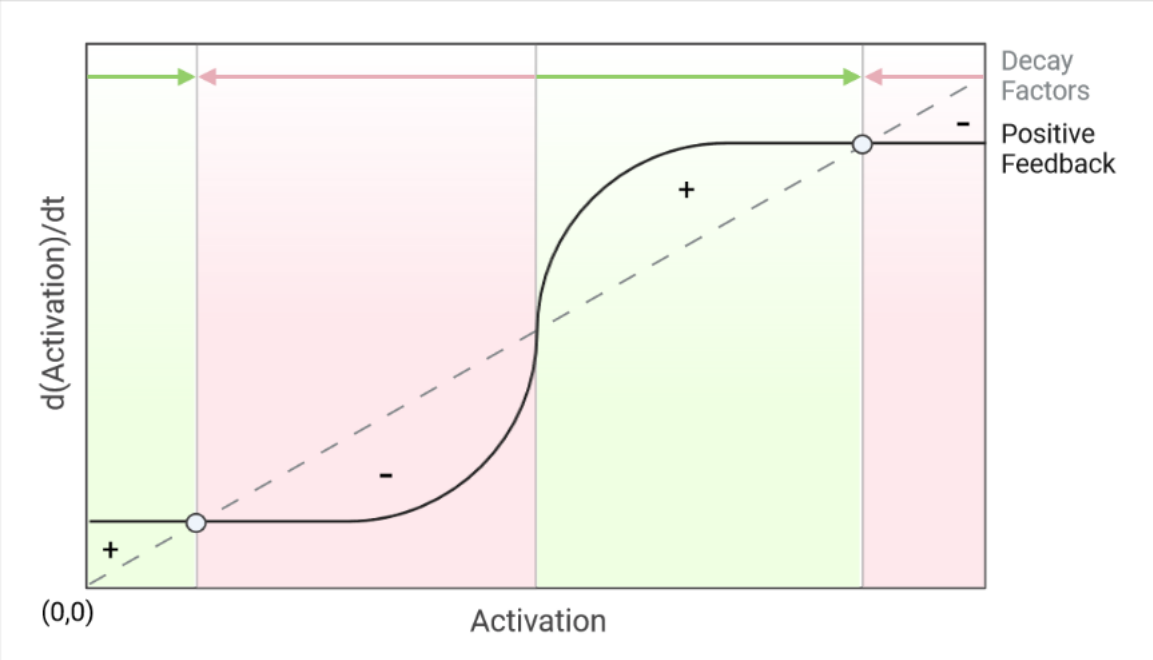

Figure 2: Descriptive illustration of the positive feedback dynamics that allow for bistability regimes. When the positive feedback-like input (solid line) overcomes the intrinsic decay factor of the system (dashed line), the system's activity is increased, and viceversa. White circles represent the attractors of the system. This "leaky switch" dynamic is found at different scales: from the molecular (Lisman, 1985) to the neural population level (Durstewitz and Seamans, 2006). Figure adapted from Chaudhuri and Fiete, 2016

.

vated by the animal's location and movement. These neurons are thought to allow the brain to retain the salient aspects of an environment which can then be translated to new settings (E. I. Moser et al., 2008). They are therefore carrying out two main functions: the formation of setting-specific place maps, which relies on mechanisms for long-term plasticity, and the generalization of previously learned experiences (in this example, features of known environments), made possible by a generative replay of past memories (M.-B. Moser et al., 2015), a process emerging from the interactions between the hippocampus and cortical areas. This replay function represents a direct link between memory and learning in the brain, and allows an efficient interplay of the two events. The biological bases of hippocampal replay are to be found in high-frequency network oscillations referred to as sharp-wave ripples. These events are characterized by specific patterns of activity over time and seem to encode for salient information about a prior behavior and its rewards, or a generated reenactment of independent experiences, useful to explore novel routes potentially leading to remembered goals (Pfeiffer, 2020). Furthermore, the number of ripple events seems to be linked to the magnitude of the reward as well as the reward prediction error associated with the replayed behavior (Roscow et al., 2021; Igata et al., 2021). This feature is useful for memory recall, navigational planning, and reward-based learning (Roscow et al., 2021; Ólafsdóttir et al., 2018). Importantly, generative replay is often inactive during the learning process, and it mainly takes place offline, i.e., during sleep (Lewis et al., 2018), or during an awake state in which no task is performed (Pfeiffer and Foster, 2013). This grants the ability to generalize salient features regardless of the current learning environment being experienced by the organism. Generative replay has recently been introduced in ANNs (van de Ven et al., 2020) and in RL agents (Roscow et al., 2021), and has been shown to aid generalization and computational efficiency of the models. The anatomical link between memory and learning is given by hippocampal interareal projections, which aid the long-term storage of memorized information. These include interactions with the cortex and connections via the previously mentioned VS (linked to MB-like representations) leading to the Ventral Tegmental Area (VTA), a region implicated in dopaminergic production (Cazé et al., 2018; Rothschild et al., 2017). Such anatomical architecture seems to reflect the need to consolidate information and integrate it in a model of the environment which

includes actual and predicted experience and reward distributions (Lowet et al., 2020). This makes the hippocampal replay buffer especially meaningful in the context of RL, given the importance of exploiting acquired knowledge about previously explored reward contingencies. Crucial functional parallels between the role of memory in the brain and in RL agents can also be drawn at the lower, "mechanistic" level of memory implementation described above. In the hippocampus, the process of generating, selecting and maintaining memories is highly dependent on oscillations in the neural networks of relevant areas, mainly sharp-wave ripples (SWR) and theta waves (prevalent during sleep) (Joo and Frank, 2018; Boyce et al., 2016). In turn, these oscillatory events are just one modality of processing, and as any other in the brain, they are sustained by differential spike timing in neural populations, which affect the connectivity of the areas. As described beforehand, this means that the consolidation or weakening of the resulting memories is essentially mediated by STDP. The coincidence of pre- and post-synaptic activity present in any network processing information, allows for the creation of a (synaptic) memory trace that can be stored and integrated within existing representations. As illustrated by equation 1, some regions of the brain pair the changes in synaptic weight resulting from spiking activity with a third factor, namely neuromodulation. Thus, three factor learning occurs with synaptic memory traces, and represents a memory function in the brain, as described by Gerstner et al., 2018:

$$\frac{d}{dt}W_{i,j} = e_{i,j}\mathrm{MOD}(t) \tag{2}$$

where MOD is the global third factor of neuromodulation, and $e_{i,j}$ is a variable set according to coincidences between pre- and post-synaptic activity. $e_{i,j}$ represents a correlation detector of the pre-synaptic neuron firing and the post-synaptic state, we previously referred to it as eligibility trace (Gerstner et al., 2018). Several biological events sustain and originate synaptic traces: from pre-synaptic voltage-gated channels and receptors to post-synaptic receptors and calcium gradients, but ultimately these events are triggered by pre- and post-synaptic spikes and are a common trait of computation in the striatum, the cortex, and the hippocampus (Fisher et al., 2017; Lim et al., 2020; Bittner et al., 2017). They allow the brain to hold data in the circuit that processes it, flagging specific information that needs to be updated. The idea of an eligibility trace that can highlight relevant information and make them modifiable was at first conceptualized in RL modeling paradigms. For instance, in some TD learning methods, only the eligible states or actions can undergo learning changes in the presence of an error (Gerstner et al., 2018). More on eligibility traces in the RL field can be found in section 2.2. Additionally, the relevant information highlighted by eligibility traces can be related to reward expectation, and be therefore influenced by neuromodulatory activity (Pan, 2005). As shown, the biological substrate of memory in the brain can offer a directly compatible method to implement eligibility traces into a three factor learning rule, tying learning and memory to the same underlying mechanisms of plasticity, and offering a suitable way to aid the efficiency of learning agents.

2.2 Learning from the brain: models of learning

2.2.1 Modeling learning modalities

The three described conceptualizations of learning (SL, UL, RL) originated from the engineering field, and for each and every one of them different methods of implementation, algorithms and architectures were developed, with a wide range of applications. SL has mainly branched into two paradigms: regression and classification. Regressors predict continuous values, while classifiers map the input space into discrete categories (Nasteski, 2017). They both have a long list of use cases, especially when implemented onto deep ANNs, including image recognition, natural language processing, and predictive analytics (Makantasis et al., 2015; Caruana and Niculescu-Mizil, 2006), but classification models are currently the ones with higher practical relevance (Sen et al., 2020). Although efforts in integrating biologically plausible features in supervised learning are present, they are mainly limited to the introduction of classic STDP-based rules in spiking neural networks (SNNs; Hao et al., 2020). For SNNs, a fundamental ceiling is given by the fact that the backpropagating instructive signal of SL cannot be directly implemented, since the neuron's spikes are non-differentiable functions. Approaches to this issue attempt at employing differentiable surrogate gradients, to enable spike-timing-dependent backpropagation (Neftci et al., 2019; Rathi et al., 2020). Other attempts with similar goals include the mixing and optimization of different learning paradigms (for instance, SL and UL; Bekolay et al., 2013). This avenue also led to the creation of hybrid paradigms like semi-supervised learning (Yang et al., 2022), the analysis of which is beyond the scope of this review. UL can be employed to carry out clustering tasks as well as dimensionality reduction, and they are also used to build generative models. The first two applications re-

late to the ability of the network to place input data points into a high dimensional feature space. Clustering algorithms group data points based on their similarity, represented by the relative position in the input space itself (Sinaga and Yang, 2020). Dimensionality reduction techniques are useful to reduce the number of features in the given dataset without losing important information about the input and the relations between its data points (Arulkumaran et al., 2017). Applications of these methods include data segmentation and visualization, as well as image and speech recognition. Generative models also learn the underlying distributions of data, but they use it to generate new samples that are similar to the original input. Application of generative models include generative adversarial networks (GANs), enhanced classifiers in which the output of a generative encoder is given as input to a decoder along with actual input data. As the encoder learns to generate more and more realistic data, the decoder will learn to discriminate it from the real input (Radford et al., 2015). In terms of biological plausibility, UL has also been successfully implemented in SNNs following STDP rules (C. Sun et al., 2022; Masquelier and Thorpe, 2007). This implementation is fairly direct since Hebb's rule depends only on local factors (pre- and post-synaptic activity) and does not require any external supervisory signal (Gerstner et al., 2014). Furthermore, UL algorithms can be designed to perform computations locally, without requiring centralized processing. This relieves computation from the SNN to redistribute it to decentralized subnetworks, a feature able to introduce parallel processing of input and to mimic the properties of spatial interaction in biological neurons (Saunders et al., 2019). Local computation can also help to create a hierarchical structure in the model, breaking down the performed task by fitting parts of it to different local sub-circuits and later merging them in a common representation (usually of lower dimensions; Lawrence K. Saul and Sam T. Roweis, 2003), a feature that is reflected at the mesoscopic level in the brain, for instance in the hierarchy of areas involved in visual processing (Hochstein and Ahissar, 2002). Interestingly, UL has been extensively used to aid other forms of learning. By learning from the structure of the data, UL can help to improve the performance of supervised models and reduce the need for large amounts of labeled data. UL can also aid RL functionality, especially when implemented with locally connected networks (Weidel et al., 2021). This avenue is particularly relevant in biologically plausible models of learning, since learning in completely new environments requires updating representations in an unsupervised fashion, which emerges from the activity of specialized cell types in different brain networks. Partitioning a SNN (specifically, its input space) to perform a specific RL task does not generalize well to realistic environments and, although it aids output performance for the desired task, it has the drawback of not reflecting the brain's functionality (Frémaux et al., 2013). Thus, letting synapses between input and representation layers in the model be sorted in an unsupervised way can be a more desirable mechanism when operating in ecological settings. Conversely, the co-expression of unsupervised Hebbian signals and reinforced (reward-based) adaptations in the same model does not come without some pitfalls in its implementation (Fremaux et al., 2010; Frémaux and Gerstner, 2016). As described in section 1, RL emerged at first as an approach to explain different aspects, from animal conditioning to control theory, and later developed into a paradigm able to manage the actions of any agent into an interactive environment. The classical applications of RL were at first related to the spatial navigation of an agent in a reward based setting, but the paradigm is often applied to a variety of cognitive tasks that can deliver iterative feedback to the learning agent. Current real-word use cases of RL include robotics, autonomous navigation, resource allocation and recommendation systems; more on these applications can be found in section 3.2. The next sections highlight the importance of integrating biologically plausible features into RL models (in this case plasticity, neuromodulation and memory replay) to expand the use cases of this paradigm both inside and outside of research.

2.2.2 Modeling plasticity and neuromodulation

As stated in section 2.1, recent literature highlights the importance of neuromodulation of Hebbian plasticity in RL, attention, and memory consolidation (Brzosko et al., 2019; Brea and Gerstner, 2016). Since SNNs compute on discrete events in time, the input is usually converted into spike trains, and encoded by differential spike timing of neurons. Other than the input population, deep SNNs (SDNN) architectures can also be employed, in which subsequent layers respond to different features of the task at hand, and do so relative to the timing with which they receive the input. Following STDP, some form of synchrony (local correlations) between populations is usually determined to be the cause of synaptic modifications. This could be directly related to the neuron's firing frequencies in more biologically detailed models (Beyeler et al., 2013; Nessler et al., 2009), or it could consider the neurons that first receive the spikes from the input population to manage

the consequent weight adjustments (usually employing integrate-and-fire neurons, with simplified rules for determining threshold potentials, as in Kheradpisheh et al., 2018; or Saunders et al., 2019). It is often the case that these networks implement some form of winner-takes-all dynamics (WTA), associating input and learned response to the strengthening of some synapses and the weakening of others. STDP in SNNs finds application in a variety of tasks. An example is image processing, in which the model employs SNNs with STDP rules in its whole architecture, including convolutional layers (in which the spiking activity is linked to feature recognition based on input from previous layers) and pooling layers (which propagate the spike pattern of feature-specific layers) (Kheradpisheh et al., 2018). Other examples include models able to carry out categorization and decision-making using Izhikevich neurons (Izhikevich, 2003), a simplified, phenomenological neuronal model comprising intrinsic adaptation and spike generation mechanisms, in which STDP is programmed to emerge from the firing behavior caused by the neuron's biophysical properties (Beyeler et al., 2013). Naturally, focusing on the paradigm of RL, a reward-linked signal is essential for the model's functionality in linking plasticity to the internal representations of environments. Besides, neuromodulatory signals are causal determiners of performance, regardless of whether they are implemented to influence representation, action policy, or other learnable elements (Xing et al., 2020; Velez and Clune, 2017). Still, merging neuromodulation with STDP offers a method of updating synaptic weights that is compatible with spiking events and simultaneously provides a biophysically inspired solution to the credit assignment problem, that usually requires bridging multiple timescales (see equation 1). For this reason, the integration of STDP and neuromodulation emerged as a logical step in the RL field, and constitutes an extremely promising avenue for research and technology development.

This "convergent evolution" of RL models is attested by the fact that the framework of three factor learning (Kuśmierz et al., 2017), with neuromodulation acting on top of some plasticity rule, is described and referred to in various contexts, albeit with some difference in the way it is implemented. Examples include neoHebbian plasticity rules (Triche et al., 2022), behavioral time scale plasticity (Gerstner et al., 2018; Bittner et al., 2017), reward-modulated Hebbian learning (Hoerzer et al., 2014; Pfeiffer, 2020), and tag-triggered consolidation (Bin Ibrahim et al., 2022; Luboeinski and Tetzlaff, 2021; Okuda et al., 2021). In all these models, the relevance of introducing neuromodulation is based on the same underlying rationale with which it is implemented in the brain (see section 2.1.2). Referring to Equation 2, a functional model that can implement three factor learning will have two basic requirements: some form of local indicator of pre- and post-synaptic correlation that can exert a flagging function over time, and a global variable for neuromodulation able to link weight changes to external inputs. Focusing on the first element, the eligibility trace $e_{i,j}$ should be a signal that enhances the strength of some synapses based on the global second variable MOD, and decays over time. This means $e_{i,j}$ should have a relatively long time constant τ_e (longer than the time constant for Hebbian modifications) that can link Hebbian plasticity to reward modulation, and should approximately correspond to the duration between starting an action and receiving a reward, i.e. should be able to solve the distal-reward problem. This can be seen as one of the core challenges of biologically plausible RL agents: determining the proper credit to be given to synaptic weights based on their role in producing positive or negative outcomes over time with spiking neurons (Triche et al., 2022). A convenient angle from which this problem can be tackled is TD learning with eligibility traces, referred to as TD(λ). In classical TD learning, the value of a given state is bootstrapped from previous rewards and from the expected utility/value of subsequent states. As mentioned in section 1, the concepts proposed by the field of dynamic programming are similar to the ones in TD learning, albeit developed for different purposes. In his work on the topic, Bellman, 1952 describe the utility of a state in which the agent is placed as the return the agent is expected to obtain if starting from that state and following a policy π:

$$V(S_t) = E_\pi[R_{t+1} + \gamma V(S_{t+1})] \tag{3}$$

Where $V(S_t)$ is the utility or value of being in the current state, $V(S_{t+1})$ is the value of the next state, γ is a temporal discounting factor (how future rewards weight on the current value) and R is the reward expected when the agent is in state S_t and takes the action prescribed by the policy, $\pi(s)$. The principle behind this equation is used in TD learning to construct predictions about subsequent temporal steps in the task, which can inform a target ($R_{t+1} + \gamma V(S_{t+1})$ in equation 4), to estimate the return of certain actions. Its simplest form, TD(0), introduces a learning rate, or step size parameter α for the update process, as described in equation 4:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{4}$$

TD(λ) introduces an eligibility trace (λ) for each state, that acts as a multiplier on the TD error in the value update equation, obtaining this general update rule from Triche et al., 2022:

$$V(S_t) \leftarrow V(S_t) + \alpha \lambda_t(S_t)[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \quad (5)$$

The value of the trace keeps track of how frequently a state has been visited and how recently these visits took place. The idea behind eligibility tracing in RL models is that states that were visited often before a reward, gain higher significance during learning, and their relevance has a limited lifespan in time. Thus, the trace itself must be updated iteratively, and different strategies are employed to capture what is essentially the effect of the time constant τ_e. The trace can be integrated via a simple update rule that governs the TD(λ) equations or, in a SNN, it can be made contingent to pre- and post-synaptic spiking (and the relative probabilities of the two events). These two approaches to update eligibility traces are commonly shared between models of STDP, but different hybrid solutions between biologically inspired and iterative updates of eligibility traces are also employed; moreover, the concept of an eligibility trace itself offers an ideal ground to implement neuromodulatory events, which will also differ depending on the model setup. The literature provides a wide span of implementations, ranging from biologically detailed models able to closely mimic experimental conditions of *in vitro* and *in vivo* circuits (Ziegler et al., 2015; Hoerzer et al., 2014) to more abstract models implementing specific bio-inspired features of modulated plasticity to perform cognitive or behavioral tasks (Miconi, 2017; Fang et al., 2021). Of course, the corresponding models will have different levels of detail depending on the objective of the study. The general function attributed to neuromodulation is often modeled directly into deep RL agents with neoHebbian learning, approximating the effects of dopamine transmission in the brain as a reward prediction error signal, and ignoring both the additional roles of dopamine in the brain, and the effects of other neuromodulatory factors. Despite the big simplifications carried by this approach, such models can still be extremely valuable in constraining hypothesis when studying the brain itself, for instance in untangling the relations between RPEs and other neuromodulatory functions. Of course, the same framework also provides significant improvements for RL agents performing in realistic environments (Fang et al., 2021; Xing et al., 2022) or in complex cognitive tasks (Miconi et al., 2020). On top of that, neuromodulators can acquire different functions, depending not only on the molecule, but also on the transitions of activity of brain regions implicated in a given behavior, and on the cell types present in those regions (Lee and Dan, 2012; Minces et al., 2017). However, research on neuromodulation as a biological process is still relatively young, and extensive work, for instance on the functionality of the receptors for these molecules, is needed to reach the level of understanding required to implement them in detailed models. Finally, another potential use of neuromodulatory concepts in modeling learning is to implement a modulatory signal without directly conferring it any biologically derived feature while allowing it to aid the performance of the network. In this case, rather then modeling the effect of any specific molecule, is the adaptive design of neuromodulatory systems that gets implemented in the network (Miconi et al., 2020 Fang et al., 2021). Miconi et al., 2020 constructed a differentiable framework in which plasticity is optimized though gradient descent, and can be applied to train and improve the performance of backpropagation models in carrying out a RL task. The author proposes the use of this emergent modulating signal to facilitate the automatic design of efficient, self-contained reinforcement learning systems, with fine tuned eligibility traces and improved reward-contingent Hebbian rules. Two versions of the model are proposed. The first one is more generally inspired by the modulation of networks (it does not model dopaminergic signalling nor eligibility traces) and has the following form:

$$\mathrm{Hebb_{i,j}}(t+1) = \mathrm{Clip}(\mathrm{Hebb_{i,j}}(t) + M(t) x_i(t-1) x_j(t)) \quad (6)$$

Here, the variable *Hebb* accumulates the product of pre- and post-synaptic activity (x_j and x_i, respectively), constrained in the interval [-1, 1] by the *Clip*(x) function. The output of the two neurons i and j is represented by the variable x in the timesteps (t) and ($t-1$), while $M(t)$ is the network-computed, time-varying neuromodulatory signal. The work also proposes a more bio-inspired version of this update rule, accounting for the known effect of dopamine and eligibility traces on Hebbian plasticity in animal brains:

$$\mathrm{Hebb_{i,j}}(t+1) = \mathrm{Clip}(\mathrm{Hebb_{i,j}}(t) + M(t) e_{i,j}(t)) \quad (7)$$

$$e_{i,j}(t+1) = (1-\eta) e_{i,j}(t) + \eta x_i(t-1) x_j(t) \quad (8)$$

Where the eligibility trace is present in the main update rule, and is updated considering an exponential average of pre- and post-synaptic activity with a decay factor η, formally equivalent to the previously mentioned time constant: $\tau_e = 1/\eta$. Considering the sparsity of data regarding the effects and dynamics

of neuromodulation in the literature, the introduction of a modulatory function that mimics the currently known features of the event in a framework compatible with STDP (or some variant of Hebbian learning) is a promising suggestion to aid models' performance as well as their ability to explain brain computation.

2.2.3 Modeling memory replay

Many of the findings regarding memory in the brain, described in section 2.1.3, sparked the motivation for a number of works investigating the computation behind memory and the relevance of hippocampal-cortical interactions in this context. Complementary Learning Systems (CLS) is a theoretical framework in neuroscience according to which the brain uses multiple, specialized learning systems to process information. Research in CLS suggests that the cortex and the hippocampus work together to handle memory transfer from short-term to long-term storage, with the hippocampus being a fast learning system, that rapidly encodes specific memories, and its interaction with the cortex constituting a slower learning system, that uses memory replay to consolidate information (Kumaran et al., 2016; O'Reilly et al., 2014; McClelland et al., 1995). CLS is therefore roughly in accordance with the anatomical studies discussed in section 2.1 (a few, more biologically detailed views on CLS can also consider other regions, as in the work of Atallah et al., 2004 where basal ganglia is included). Recent theoretical neuroscience advancements uncovered how the interplay between this fast and slow learning systems through memory replay might in fact be covering the function of optimizing generalization. The amount of consolidation though memory replay from hippocampus to cortex is also found to be dependent on the level of predictability of the environment that is being experienced (W. Sun et al., 2023). As stated previously, the principles behind memory replay and reinforcement mechanisms are highly compatible, at least in the brain. For this reason, studies successfully attempted to implement CLS in TD learning, as in the work of Blakeman and Mareschal, 2020, using deep Q-learning networks (DQNs). Following a similar route, W. Sun et al., 2023 investigated the theoretical framework of CLS by modeling it into a Hebbian learning paradigm. From a Machine Learning (ML) point of view, the necessity for this new theoretical framework comes from the fact that "blindly" transferring all data from a specialized fast memory unit like the hippocampus to a generalization structure such as the cortex would not allow the learning agent to adapt to realistic settings. Looking at the brain through ML lenses, this problem would not present itself in noise-free environments, since the cortex offers a huge parameter space and can hardly over-fit (Hasson et al., 2020), but becomes relevant when noise is present and for instance, reward inputs are sparse. Uninterrupted learning from the same batch of data can cause learning agents to pick up spurious correlation in those data, which means that hippocampal replay has to be a moderated process that allows to pick up general rules of different lived experiences. Hence, a description of the modalities of this information transfer, and ultimately of the relationship between fast and slow memory processing in learning is required, and especially so in noisy, realistic environments. Conceptually, both in brains and machines, the interplay of fast and slow learning systems is crucial to prevent one of the main issues of RL described in section 1, namely, catastrophic interference. In this case, the fast system would mainly record information from the current environment, which would then be used to train the slower learning system to consolidate generalized representations (Botvinick et al., 2019). This is then one way in which fairly direct inspiration from the biology of the brain can give significant insights into how the big obstacles in RL can be tackled. The last remaining step to achieve biologically plausible memory consolidation is to include the offline replay function in models, recent strategies at implementing it can be found in the work of van de Ven et al., 2020 and Barry and Love, 2022, for a review, see Roscow et al., 2021. Lastly, referring back to the findings of W. Sun et al., 2023, consolidation via memory replay is postulated to be dependent on the inferred predictability of an environment. This means that the agent would assess through experience the degree to which events in that environment are predictable, and adjust memory consolidation based on that. Although the assessment itself is subject-dependent and often has various evolutionary constraints, the mechanism that implements it is well known to be neuromodulation. In fact, if the subject experiences reward events, this assessment would just be computing RPEs, a process that is already well established in RL. These proofs of concept, at the frontier between theoretical and computational neuroscience, demonstrate how a neoHebbian account on learning, complete with a description of eligibility traces (which, as highlighted in section 2.2.2, significantly aids models performance in realistic environments by modulating reward contingencies) can be compatible with the theories of memory replay (which is believed to aid generalization of the learned reward contingencies) by using common synaptic update rules. A logical next step in this field of research seems to be capitalizing on this compatibility, by

constructing RL agents that, operating via spike timing, can support three factor learning, sustained activity, as well as an efficient replay buffer that can use these two principles to optimize generalization.

3 Conclusions

3.1 Discussion

There seem to be two main knowledge gaps impeding the development of efficient bio-plausible RL models able to reflect brain computation. The first gap is the previously mentioned sparsity of data around neuromodulation, see section 2.2.2. Our knowledge of neuromodulatory functions is still confined to a description of the general effects of each neuromodulator on a large scale, but not enough is known about their activity on a network scale and the results of interaction between more neuromodulators. As a result, in RL, RPE is considered when modeling neuromodulatory processes, without taking into account that dopamine, which implements RPEs in the brain, seems to have a variety of additional functions in different areas and cell types. A few exceptions are starting to prove the validity of integrating more realistic and diverse neuromodulatory functions to aid RL models' performance (Xing et al., 2022; Zannone et al., 2018), or to advance our understanding of these processes in the brain (Graupner and Gutkin, 2009). The second gap revolves around determining the best learning rule to apply to these RL agents. In fact, eligibility traces do represent a nice bridge between STDP and reward contingent modulation, but only under the assumption that they can always cover the temporal gap between action and reward. A number of studies suggest that the brain can learn reward contingencies even with substantial delays, and once again, the timing of these plasticity rules seem to vary between regions and cell types (Suvrathan, 2019). Although the concept of plasticity being driven by spike timing can represent a good common mechanism for learning, a model that is able to operate with three factor learning and sustain an offline memory replay might require differential time-scales for its eligibility traces, for instance depending on the function it is currently carrying out. This approach is not unexplored in the Deep RL field, since parameter dependencies in deep models constitute an issue when we want to maintain a fixed eligibility trace (Kobayashi, 2022). The possibility of expanding on STDP rules was also explored in SNNs performing RL, since plasticity linked to a fixed spike timing dependency seems to be a constraining factor when operating in dynamic settings (Jimenez Rezende and Gerstner, 2014). Of course, the current limits of knowledge set by experimental neuroscience do not represent a valid reason to give up advancing our understanding of RL in the brain and in models. When it comes to deep RL with neoHebbian learning, the main improvement given by this implementation revolves around the ability to face sparse reward, dynamic environments. While the known mechanistic elements for a neoHebbian principle are relatively well characterized in models, it could be possible that further improvements can be made in the fuzzier, more delicate realm of representations. An example is given by the work of Ben-Iwhiwhu et al., 2022, in which neuromodulation acts on a meta-RL network that can inform better policies in complex environments and enrich the agent's representations. Furthermore, three factor learning is still a global framework mainly directed towards reward optimization, but more can be accounted for in the interplay of short- and long-term rewards, which seems to have a big influence in forming representations, creating complex action plans, and aiding generalization. The way to implement this more realistic rewards and its integration in RL models is of course an open area of debate, since one could approach the problem either from a deep RL perspective, proposing new architectures able to sustain this differential rewards, but a potentially valid alternative can also be to integrate new (possibly bio-inspired) systems that can account for more than just reward to increase depth in the agent's decision-making (Silver et al., 2021; Vamplew et al., 2022). Deep RL also sees exciting new obstacles in the interplay of learning modalities. As mentioned in section 2.2, neoHebbian RL already mixes UL aspects to RL (not without challenges). Explicitly expanding on this integration could lead to obtaining agents able to perform learning in a broader sense of the term, that would have reward maximization as just one facet of their capabilities (Weidel et al., 2021). An ulterior interesting addition would be to consider the constraints given by evolution described in section 2.1.1, which could inform prior inductive biases about the natural environment around us and the way we perceive it. A theoretical framework that considers the evolutionary learning component while still accounting for the shortcomings of a single reward maximization function is homeostatic learning (Keramati and Gutkin, 2014). In this perspective, the reinforcement signal is not limited to rewards or predictions about reward, but to the maintenance of a set of homeostatic values in the face of changing environments, updating a temporal discount of physiological variables (Laurençon et al., 2021; de Abril and Kanai, 2018). Note that the idea of homeostatic learning can be envisioned as an alternative or a complement to the other approaches we refer. In sum, regarding future

outlooks for this field of research, it seems more and more important to highlight the need for theoretical approaches to shine light on new potential avenues in computational neuroscience, which could contribute to models performing better in realistic environments (aiding technology applications), as well as being able to uncover new aspects of brain computation. As mentioned in this review and postulated by several studies, learning seems to involve at least two speeds. At the microscopic, mechanistic scale, fast synaptic transmission is accompanied by the slower event of neuromodulation. In the mesoscopic, functional realm, the human brain exploits two systems of learning: a fast hippocampus-dependent process which fine tunes representations, and a slow cortical-dependent learning which optimizes generalization and allows us to form a coherent view of the world (Botvinick et al., 2019). Both fast and slow processes are fundamentally important, and a functional brain is only possible with the interaction of the two. Taking such interplay as a metaphor, we should learn from the brain, and not discard the slower proceedings of theoretical neuroscience in favour of a fast paced, data-driven fine tuning of currently existing models. It is true that fast incremental progress driven by advancing technologies can lead to explore novel grounds for research, and indeed there is great value in implementing current known biological features in RL models to aid their performance, but the slower efforts in understanding the theoretical principles of neural computation are what allow us to give direction and meaning to this new frontiers, ultimately benefiting human development.

3.2 Interdisciplinary Reflections

3.2.1 Neuromorphic Computing Applications

The idea of imitating biological neurons by fusing together the two elements of the von Neumann architecture (memory and processing) is not only interesting for the study of the brain, but also to aid model's computational efficiency. This problem arises from the high energy consumption of fetching massive amounts of data from a memory unit to a processing network. Holding memory in each processing neuron in the network would greatly aid energy efficiency and computational costs, a possibility that led to the creation of neuromorphic hardware able to support this new bio-inspired framework. An example of this framework is given by the memristors, transistors that function as an analogue valve of information (current) instead of as a switch with a definite threshold. These can represent artificial synapses that govern information flow via current resistance (Thomas, 2013). Of course, implementing all of these elements into a densely connected network on a dedicated hardware is a huge challenge, in the face of which different variations of this synaptic structure were proposed. The structure of memristors offers itself nicely for the use of SNNs, since it can more easily account for spike timing. The main use case for memristive devices is unsupervised learning following STDP rules (Bill and Legenstein, 2014). However, reinforcement learning can also be implemented along with the WTA dynamics of the SNN (Mehonic et al., 2020). As stated, a non-von Neumann architecture is especially useful for computational efficiency. for instance if we want to introduce reward-modulated spike-timing dependent plasticity (R-STDP) (Shi et al., 2021; Wunderlich et al., 2019). The system also potentially allows to include a memory replay function. The need for innovative ways to implement high number of synaptic connections and communicating networks into a neuromorphic hardware seems to be the current main obstacle to adapt more complete bio-plausible RL frameworks into neuromorphic computing. Candidate solution to face it include implementing oscillator-based systems, in which synapses are represented by oscillating waves of current and their strengthening and weakening is given by synchronization events between these waves, the pattern of which could encode for a specific stimulus or internal representation (Romera et al., 2018).

3.2.2 Studying the brain with targeted manipulations

Brain inspired models are getting more and more effective in improving energy efficiency, but even without considering the limitations of computational costs, the problem of how to relate back these models to explain brain computation remains non-trivial. If computational cost is on one end of the scale, the degree of biological detail sits on the other, and the trade-off needs to be adjusted according to the research question at hand. A vivid example of the applications of biologically plausible RL models to uncover brain computation can be found in the study of neuromodulators (see section 2.2.2). In this field, different models are employed at different levels of description to assess how their local effects on neural populations influences global connectivity and ultimately, behavior. The functional nature of neuromodulation still needs to be uncovered, and behavioral neuroscience can provide significant data on the interplay of these factors by silencing a specific type of neuromodulator, or its effect on a given region. This data can then be confronted with ad-hoc build RL models, which can potentially allow for a multi-scale comparison with *in vivo* behavioral experiments, and can be exploited as powerful explana-

tory tools. In a future outlook, modelling the effects of a single neuromodulator could aid crucial advancements in understanding the roles of their interactions in the human brain, as proposed by Mei et al., 2022. Or, in more physiologically detailed models, we could maintain the same neuromodulatory process across different brain regions and cell types, to assess the degree of relevant variability in that scenario. Furthermore, extending the role of the functional study of neuromodulators, a bio-realistic model could help in understanding the effects of neurodegenerative diseases (processes in which neuromodulation can be significantly compromised) like Parkinson's disease (PD) (Liebenow et al., 2022), or it can help to uncover the neurobiological mechanisms of less explored conditions, like Anhedonia (Kangas et al., 2022).

3.2.3 Robotics

A natural application to examine RL models that face dynamic environments includes autonomous agents able to interact with these environments successfully, achieving desired goals. This impulse pushed toward developments in the robotics field, in which a few complications to the employment of functional ML algorithms (and RL agents) need to be considered. Firstly, practical challenges are to be found in the implementation of efficient learning agents in the robot's hardware, in a trade-off between robustness towards real-world settings and processing capabilities (Prorok et al., 2021). Emerging resilient behavior towards ecological settings can also be generated by achieving complex reward structures or relying on homeostatic principles, as discussed in section 3.1. Furthermore, the RL agent that must now be embodied in a physical structure has to gather data from sensory input units. This means sensory modules in the robot are progressively generating data from lived experiences rather than extracting data from an existing dataset. On top of that, robots are supposed to manipulate the environment around them, causing dynamic changes to it, to which the robot itself must adapt when carrying out complex tasks. These can be considered obstacles because the raw sensory input must be contextualized by the robot itself in a coherent representation of the task, held as a part of an explicit model of the world, over which behavioral planning and control must be added to cause an action. In turn, the robot acting on its surroundings updates this whole chain of events continuously and therefore requires constant adaptation to these changes (Hernández et al., 2018). All these challenges would benefit from a system operating within sparse data, dynamic settings, ultimately able to address both the problem of optimizing inductive bias and preventing catastrophic interference. RL agents developed in an SNN or a deep network able to include neuromodulation and memory replay could represent good candidates to face these two main issues. The robot would also need to be able to operate in a non-supervised fashion, either defining correct rewards for the desired task or following the right reinforced signals (for instance, a measure of closeness to predictions). Creating an end-to-end learning robot would demand the interplay of different learning modalities, a requirement that is very hardly accomplished without taking inspiration from biological brains. All of the central themes described in this review are compatible with these needs: biologically plausible deep RL agents could help robotics in acquiring more complex and realistic loss functions to navigate the real world, even though the technical problems of running such framework in a performing hardware cannot be ignored. Conceptually, catastrophic interference and inductive biases are still the two main challenges in the field, since robotics aims at providing agents able to navigate complex and dynamic settings. Meta-learning as a framework proposes to face the agent with a distribution of tasks, with the underlying goal of exploiting their commonalities to aid its performance by constructing effective priors. The process is led by the ability of the agent to generalize across different environments, and its improvement can bring clear advantages in the field (Clavera et al., 2018; Finn et al., 2017). Having realistic priors would greatly aid human-machine interaction since effective cooperation between the two is enhanced by having similar inductive biases, which would give rise to more easily explainable behavior. Meta-learning algorithms can be implemented in RL agents and aided with human-like representations (Arndt et al., 2020). Approaches to implement it also include training specific architectures to learn human-like prior knowledge about an environment and inform an existing RL agent, in an attempt to approach the function of human episodic memory (Ritter et al., 2018). Alternatively, realistic priors can be extracted from input data, in case information about human priors on the environment is available (Kumar et al., 2022). In our case, the memory replay unit would aid the machines in learning natural priors, while efficient memory storage can help to prevent catastrophic interference.

References

Anisimova, M., van Bommel, B., Wang, R., Mikhaylova, M., Wiegert, J. S., Oertner, T. G., & Gee, C. E. (2022). Spike-timing-dependent plasticity rewards synchrony rather than causality. *Cerebral Cortex*, *33*(1), 23–34. https://doi.org/10.1093/cercor/bhac050

Arndt, K., Hazara, M., Ghadirzadeh, A., & Kyrki, V. (2020). Meta Reinforcement Learning for Sim-to-real Domain Adaptation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2725–2731. https://doi.org/10.1109/ICRA40945.2020.9196540

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, *34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240

Atallah, H. E., Frank, M. J., & O'Reilly, R. C. (2004). Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. *Neurobiology of Learning and Memory*, *82*(3), 253–267. https://doi.org/10.1016/j.nlm.2004.06.004

Averbeck, B., & O'Doherty, J. P. (2022). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*(1), 147–162. https://doi.org/10.1038/s41386-021-01108-0

Badre, D., & Frank, M. J. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico–Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536. https://doi.org/10.1093/CERCOR/BHR117

Barry, D. N., & Love, B. C. (2022). A neural network account of memory replay and knowledge consolidation. *Cerebral Cortex*, *33*(1), 83–95. https://doi.org/10.1093/cercor/bhac054

Bekolay, T., Kolbeck, C., & Eliasmith, C. (2013). Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*. https://escholarship.org/uc/item/9k64b389

Bellman, R. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences*, *38*(8). https://doi.org/10.1073/pnas.38.8.716

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Ben-Iwhiwhu, E., Dick, J., Ketz, N. A., Pilly, P. K., & Soltoggio, A. (2022). Context meta-reinforcement learning via neuromodulation. *Neural Networks*, *152*, 70–79. https://doi.org/10.1016/j.neunet.2022.04.003

Beyeler, M., Dutt, N. D., & Krichmar, J. L. (2013). Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule. *Neural Networks*, *48*, 109–124. https://doi.org/10.1016/j.neunet.2013.07.012

Bill, J., & Legenstein, R. (2014). A compound memristive synapse model for statistical learning through STDP in spiking neural networks. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00412

Bin Ibrahim, M. Z., Benoy, A., & Sajikumar, S. (2022). Long-term plasticity in the hippocampus: maintaining within and 'tagging' between synapses. *The FEBS Journal*, *289*(8), 2176–2201. https://doi.org/10.1111/febs.16065

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., & Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science*, *357*(6355), 1033–1036. https://doi.org/10.1126/science.aan3846

Björklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, *30*(5), 194–202. https://doi.org/10.1016/j.tins.2007.03.006

Blakeman, S., & Mareschal, D. (2020). A complementary learning systems approach to temporal difference learning. *Neural Networks*, *122*, 218–230. https://doi.org/10.1016/j.neunet.2019.10.011

Bliss, T. V. P., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of Physiology*, *232*(2), 331–356. https://doi.org/10.1113/jphysiol.1973.sp010273

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron*, *107*(4), 603–616. https://doi.org/10.1016/j.neuron.2020.06.014

Boyce, R., Glasgow, S. D., Williams, S., & Adamantidis, A. (2016). Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science*, *352*(6287), 812–816. https://doi.org/10.1126/science.aad5252

Brea, J., & Gerstner, W. (2016). Does computational neuroscience need new synaptic learning paradigms? *Current Opinion in Behavioral Sciences*, *11*, 61–66. https://doi.org/10.1016/j.cobeha.2016.05.012

Bremner, J. G., Slater, A. M., & Johnson, S. P. (2015). Perception of Object Persistence: The Origins of Object Permanence in Infancy. *Child Development Perspectives*, *9*(1), 7–13. https://doi.org/10.1111/CDEP.12098

Brzosko, Z., Mierau, S. B., & Paulsen, O. (2019). Neuromodulation of Spike-Timing-Dependent Plasticity: Past, Present, and Future. *Neuron*, *103*(4), 563–581. https://doi.org/10.1016/j.neuron.2019.05.041

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The Human Hippocampus and Spatial and Episodic Memory. *Neuron*, *35*(4), 625–641. https://doi.org/10.1016/S0896-6273(02)00830-9

Calisir, S., & Pehlivanoglu, M. K. (2019). Model-Free Reinforcement Learning Algorithms: A Survey. *2019 27th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU.2019.8806389

Caporale, N., & Dan, Y. (2008). Spike Timing-Dependent Plasticity: A Hebbian Learning Rule. https://doi.org/10.1146/annurev.neuro.31.060407.125639

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 161–168. https://doi.org/10.1145/1143844.1143865

Cazé, R., Khamassi, M., Aubin, L., & Girard, B. (2018). Hippocampal replays under the scrutiny of reinforcement learning models. *Journal of Neurophysiology*, *120*(6), 2877–2896. https://doi.org/10.1152/jn.00145.2018

Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience*, *19*(3), 394–403. https://doi.org/10.1038/nn.4237

Clavera, I., Nagabandi, A., Fearing, R. S., Abbeel, P., Levine, S., & Finn, C. (2018). Learning to Adapt: Meta-Learning for Model-Based Control. *CoRR*, *abs/1803.11347*. http://arxiv.org/abs/1803.11347

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Costa, V. D., & Averbeck, B. B. (2020). Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore–Exploit Tradeoffs. *The Journal of Neuroscience*, *40*(12), 2553–2561. https://doi.org/10.1523/JNEUROSCI.2355-19.2020

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. https://doi.org/10.1016/j.conb.2008.08.003

de Abril, I. M., & Kanai, R. (2018). Curiosity-Driven Reinforcement Learning with Homeostatic Regulation. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–6. https://doi.org/10.1109/IJCNN.2018.8489075

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., Dolan, R. J., Heinz, A., & Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*(5), 1595–1600. https://doi.org/10.1073/pnas.1417219112

Dixon, M. L., & Dweck, C. S. (2022). The amygdala and the prefrontal cortex: The co-construction of intelligent decision-making. *Psychological Review*, *129*(6), 1414–1441. https://doi.org/10.1037/rev0000339

Doyon, J., Song, A. W., Karni, A., Lalonde, F., Adams, M. M., & Ungerleider, L. G. (2002). Experience-dependent changes in cerebellar contributions to motor sequence learning. *Proceedings of the National Academy of Sciences*, *99*(2), 1017–1022. https://doi.org/10.1073/pnas.022615199

Duff, M. C., Covington, N. V., Hilverman, C., & Cohen, N. J. (2020). Semantic Memory and the Hippocampus: Revisiting, Reaffirming, and Extending the Reach of Their Critical Relationship. *Frontiers in Human Neuroscience*, *13*. https://doi.org/10.3389/fnhum.2019.00471

Durstewitz, D., & Seamans, J. (2006). Beyond bistability: Biophysics and temporal dynamics of working memory. *Neuroscience*, *139*(1), 119–133. https://doi.org/10.1016/j.neuroscience.2005.06.094

Fang, H., Zeng, Y., & Zhao, F. (2021). Brain Inspired Sequences Production by Spiking Neural Networks With Reward-Modulated STDP. *Frontiers in Computational Neuroscience*, *15*. https://doi.org/10.3389/fncom.2021.612041

Filimon, F., Nelson, J. D., Sejnowski, T. J., Sereno, M. I., & Cottrell, G. W. (2020). The ventral striatum dissociates information expectation, reward anticipation, and reward receipt. *Proceedings of the National Academy of Sciences*, *117*(26), 15200–15208. https://doi.org/10.1073/pnas.1911778117

Findling, C., Chopin, N., & Koechlin, E. (2020). Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, *5*(1), 99–112. https://doi.org/10.1038/s41562-020-00971-z

Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 1126–1135). PMLR. https://proceedings.mlr.press/v70/finn17a.html

Fisher, S. D., Robertson, P. B., Black, M. J., Redgrave, P., Sagar, M. A., Abraham, W. C., & Reynolds, J. N. (2017). Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nature Communications*, *8*(1), 334. https://doi.org/10.1038/s41467-017-00394-x

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Fremaux, N., Sprekeler, H., & Gerstner, W. (2010). Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity. *Journal of Neuroscience*, *30*(40), 13326–13337. https://doi.org/10.1523/JNEUROSCI.6249-09.2010

Frémaux, N., & Gerstner, W. (2016). Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Frontiers in Neural Circuits*, *9*. https://doi.org/10.3389/fncir.2015.00085

Frémaux, N., Sprekeler, H., & Gerstner, W. (2013). Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons. *PLoS Computational Biology*, *9*(4), e1003024. https://doi.org/10.1371/journal.pcbi.1003024

Gallistel, C. R., & King, A. P. (2009). *Memory and the Computational Brain*. Wiley. https://doi.org/10.1002/9781444310498

Geerts, J. P., Chersi, F., Stachenfeld, K. L., & Burgess, N. (2020). A general model of hippocampal and dorsal striatal learning and decision making. *Proceedings of the National Academy of Sciences*, *117*(49), 31427–31437. https://doi.org/10.1073/pnas.2007981117

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal Dynamics*. Cambridge University Press. https://doi.org/10.1017/CBO9781107447615

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*. https://doi.org/10.3389/fncir.2018.00053

Giese, K. P., Fedorov, N. B., Filipkowski, R. K., & Silva, A. J. (1998). Autophosphorylation at Thr ¡sup¿286¡/sup¿ of the α Calcium-Calmodulin Kinase II in LTP and Learning. *Science*, *279*(5352), 870–873. https://doi.org/10.1126/science.279.5352.870

Graupner, M., & Gutkin, B. (2009). Modeling nicotinic neuromodulation from global functional and network levels to nAChR based mechanisms. *Acta Pharmacologica Sinica*, *30*(6), 681–693. https://doi.org/10.1038/aps.2009.87

Groman, S. M., Keistler, C., Keip, A. J., Hammarlund, E., DiLeone, R. J., Pittenger, C., Lee, D., & Taylor, J. R. (2019). Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. *Neuron*, *103*(4), 734–746. https://doi.org/10.1016/j.neuron.2019.05.042

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. https://doi.org/10.1016/j.neunet.2019.09.007

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/J.NEURON.2019.12.002

Hebb, D. O. (1950). The organization of behavior: A neuropsychological theory. New York: John Wiley and Sons, Inc., 1949. *Science Education*, *34*(5), 336–337. https://doi.org/10.1002/sce.37303405110

Hernández, C., Bermejo-Alonso, J., & Sanz, R. (2018). A self-adaptation framework based on functional knowledge for augmented autonomy in robots. *Integrated Computer-Aided Engineering*, *25*, 157–172. https://doi.org/10.3233/ICA-180565

Hochstein, S., & Ahissar, M. (2002). View from the Top. *Neuron*, *36*(5), 791–804. https://doi.org/10.1016/S0896-6273(02)01091-7

Hoerzer, G. M., Legenstein, R., & Maass, W. (2014). Emergence of Complex Computational Structures From Chaotic Neural Networks Through Reward-Modulated Hebbian Learning. *Cerebral Cortex*, *24*(3), 677–690. https://doi.org/10.1093/cercor/bhs348

Igata, H., Ikegaya, Y., & Sasaki, T. (2021). Prioritized experience replays on a hippocampal predictive map for learning. *Proceedings of the National Academy of Sciences*, *118*(1). https://doi.org/10.1073/pnas.2011266118

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572. https://doi.org/10.1109/TNN.2003.820440

Jessup, R. K., & O'Doherty, J. P. (2011). Human Dorsal Striatal Activity during Choice Discriminates Reinforcement Learning Behavior from the Gambler's Fallacy. *Journal of Neuroscience*, *31*(17), 6296–6304. https://doi.org/10.1523/JNEUROSCI.6421-10.2011

Jimenez Rezende, D., & Gerstner, W. (2014). Stochastic variational learning in recurrent spiking networks. *Frontiers in Computational Neuroscience*, *8*. https://doi.org/10.3389/fncom.2014.00038

Joo, H. R., & Frank, L. M. (2018). The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience*, *19*(12), 744–757. https://doi.org/10.1038/s41583-018-0077-1

Kangas, B. D., Der-Avakian, A., & Pizzagalli, D. A. (2022). Probabilistic Reinforcement Learning and Anhedonia. https://doi.org/10.1007/7854_2022_349

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, *3*. https://doi.org/10.7554/eLife.04811

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., & Masquelier, T. (2018). STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, *99*, 56–67. https://doi.org/10.1016/j.neunet.2017.12.005

Kobayashi, T. (2022). Adaptive and multiple time-scale eligibility traces for online deep reinforcement learning. *Robotics and Autonomous Systems*, *151*, 104019. https://doi.org/10.1016/j.robot.2021.104019

Kohonen, T. (2001). *Self-Organizing Maps* (Vol. 30). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56927-2

Krakauer, J. W., & Shadmehr, R. (2006). Consolidation of motor memory. *Trends in Neurosciences*, *29*(1), 58–64. https://doi.org/10.1016/j.tins.2005.10.003

Kumar, S., Correa, C., Dasgupta, I., Marjieh, R., Hu, M., Hawkins, R., Daw, N., Cohen, J., Narasimhan, K., & Griffiths, T. (2022). Using Natural Language and Program Abstractions to Instill Human Inductive Biases in Machines. https://doi.org/10.48550/arXiv.2205.11558

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/J.TICS.2016.05.004

Kuśmierz, Ł., Isomura, T., & Toyoizumi, T. (2017). Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology*, *46*, 170–177. https://doi.org/10.1016/j.conb.2017.08.020

Laurençon, H., Ségerie, C.-R., Lussange, J., & Gutkin, B. S. (2021). Continuous Homeostatic Reinforcement Learning for Self-Regulated Autonomous Agents. https://doi.org/https://doi.org/10.48550/arXiv.2109.06580

Lawrence K. Saul & Sam T. Roweis. (2003). Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. *Journal of Machine Learning Research*, *4*, 119–155. https://dl.acm.org/doi/10.1162/153244304322972667

Lee, S.-H., & Dan, Y. (2012). Neuromodulation of Brain States. *Neuron*, *76*(1), 209–222. https://doi.org/10.1016/j.neuron.2012.09.012

Lewis, P. A., Knoblich, G., & Poe, G. (2018). How Memory Replay in Sleep Boosts Creative Problem-Solving. *Trends in Cognitive Sciences*, *22*(6), 491–503. https://doi.org/10.1016/j.tics.2018.03.009

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00828

Liebenow, B., Jones, R., DiMarco, E., Trattner, J. D., Humphries, J., Sands, L. P., Spry, K. P., Johnson, C. K., Farkas, E. B., Jiang, A., & Kishida, K. T. (2022). Computational reinforcement learning, reward (and punishment), and dopamine in psychiatric disorders. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.886297

Lim, D.-H., Yoon, Y. J., Her, E., Huh, S., & Jung, M. W. (2020). Active maintenance of eligibility trace in rodent prefrontal cortex. *Scientific Reports*, *10*(1), 18860. https://doi.org/10.1038/s41598-020-75820-0

Lisman, J. E. (1985). A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. *Proceedings of the National Academy of Sciences*, *82*(9), 3055–3057. https://doi.org/10.1073/pnas.82.9.3055

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., & Uchida, N. (2020). Distributional Reinforcement Learning in the Brain. *Trends in Neurosciences*, *43*(12), 980–997. https://doi.org/10.1016/J.TINS.2020.09.004

Luboeinski, J., & Tetzlaff, C. (2021). Memory consolidation and improvement by synaptic tagging and capture in recurrent neural networks. *Communications Biology*, *4*(1), 275. https://doi.org/10.1038/s42003-021-01778-y

Makantasis, K., Karantzalos, K., Doulamis, A., & Doulamis, N. (2015). Deep supervised learning for hyperspectral data classification through convolutional neural networks. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4959–4962. https://doi.org/10.1109/IGARSS.2015.7326945

Masquelier, T., & Thorpe, S. J. (2007). Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. *PLoS Computational Biology*, *3*(2), e31. https://doi.org/10.1371/journal.pcbi.0030031

Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., Arai, R., & Kaneko, T. (2009). Single Nigrostriatal Dopaminergic Neurons Form Widely Spread and Highly Dense Axonal Arborizations in the Neostriatum. *Journal of Neuroscience*, *29*(2), 444–453. https://doi.org/10.1523/JNEUROSCI.4029-08.2009

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. *Journal of Neuroscience*, *31*(7), 2700–2705. https://doi.org/10.1523/JNEUROSCI.5499-10.2011

Mehonic, A., Sebastian, A., Rajendran, B., Simeone, O., Vasilaki, E., & Kenyon, A. J. (2020). Memristors—From In-Memory Computing, Deep Learning Acceleration, and Spiking Neural Networks to the Future of Neuromorphic and Bio-Inspired Computing. *Advanced Intelligent Systems*, *2*(11), 2000085. https://doi.org/10.1002/aisy.202000085

Mei, J., Muller, E., & Ramaswamy, S. (2022). Informing deep neural networks by multiscale principles of neuromodulatory systems. *Trends in Neurosciences*, *45*(3), 237–250. https://doi.org/10.1016/j.tins.2021.12.008

Miconi, T. (2017). Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, *6*. https://doi.org/10.7554/eLife.20899

Miconi, T., Rawal, A., Clune, J., & Stanley, K. O. (2020). Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity. https://doi.org/https://doi.org/10.48550/arXiv.2002.10585

Minces, V., Pinto, L., Dan, Y., & Chiba, A. A. (2017). Cholinergic shaping of neural correlations. *Proceedings of the National Academy of Sciences*, *114*(22), 5725–5730. https://doi.org/10.1073/pnas.1621493114

Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2023). Model-based Reinforcement Learning: A Survey. *Foundations and Trends® in Machine Learning*, *16*(1), 1–118. https://doi.org/10.1561/2200000086

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. *Annual Review of Neuroscience*, *31*(1), 69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723

Moser, M.-B., Rowland, D. C., & Moser, E. I. (2015). Place Cells, Grid Cells, and Memory. *Cold Spring Harbor Perspectives in Biology*, *7*(2), a021808. https://doi.org/10.1101/cshperspect.a021808

Nabavi, S., Fox, R., Proulx, C. D., Lin, J. Y., Tsien, R. Y., & Malinow, R. (2014). Engineering a memory with LTD and LTP. *Nature*, *511*(7509), 348–352. https://doi.org/10.1038/nature13294

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks. *IEEE Signal Processing Magazine*, *36*(6), 51–63. https://doi.org/10.1109/MSP.2019.2931595

Nessler, B., Pfeiffer, M., & Maass, W. (2009). STDP enables spiking neurons to detect hidden causes of their inputs. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/a5cdd4aa0048b187f7182f1b9ce7a6a7-Paper.pdf

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100. https://doi.org/10.1016/J.COBEHA.2014.10.004

Okuda, K., Højgaard, K., Privitera, L., Bayraktar, G., & Takeuchi, T. (2021). Initial memory consolidation and the synaptic tagging and capture hypothesis. *European Journal of Neuroscience*, *54*(8), 6826–6849. https://doi.org/10.1111/ejn.14902

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology*, *28*(1), R37–R50. https://doi.org/10.1016/j.cub.2017.10.073

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary Learning Systems. *Cognitive Science*, *38*(6), 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Pan, W.-X. (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. https://doi.org/10.1523/JNEUROSCI.1478-05.2005

Pavlov, I. P. (1928). *Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous activity (behaviour) of animals.* Liverwright Publishing Corporation. https://doi.org/10.1037/11081-000

Pawlak, V. (2010). Timing is not everything: neuromodulation opens the STDP gate. *Frontiers in Synaptic Neuroscience*, *2*. https://doi.org/10.3389/fnsyn.2010.00146

Pfeiffer, B. E. (2020). The content of hippocampal "replay". *Hippocampus*, *30*(1), 6–18. https://doi.org/10.1002/hipo.22824

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74–79. https://doi.org/10.1038/nature12112

Prorok, A., Malencia, M., Carlone, L., Sukhatme, G., Sadler, B. M., & Kumar, V. (2021). Beyond Robustness: A Taxonomy of Approaches towards Resilient Multi-Robot Systems.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.

Rathi, N., Srinivasan, G., Panda, P., & Roy, K. (2020). Enabling Deep Spiking Neural Networks with Hybrid Conversion and Spike Timing Dependent Backpropagation.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning: Current Research and Theory.*

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been There, Done That: Meta-Learning with Episodic Recall. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 4354–4363). PMLR. https://proceedings.mlr.press/v80/ritter18a.html

Rodhom, C., & Tolman, E. C. (1950). Purposive Behavior in Animals and Men. *The American Journal of Psychology*, *63*(2). https://doi.org/10.2307/1418946

Romera, M., Talatchian, P., Tsunegi, S., Abreu Araujo, F., Cros, V., Bortolotti, P., Trastoy, J., Yakushiji, K., Fukushima, A., Kubota, H., Yuasa, S., Ernoult, M., Vodenicarevic, D., Hirtzlin, T., Locatelli, N., Querlioz, D., & Grollier, J. (2018). Vowel recognition with four coupled spin-torque nano-oscillators. *Nature*, *563*(7730), 230–234. https://doi.org/10.1038/s41586-018-0632-y

Roscow, E. L., Chua, R., Costa, R. P., Jones, M. W., & Lepora, N. (2021). Learning offline: memory replay in biological and artificial reinforcement learning. *Trends in Neurosciences*, *44*(10), 808–821. https://doi.org/10.1016/j.tins.2021.07.007

Rothschild, G., Eban, E., & Frank, L. M. (2017). A cortical–hippocampal–cortical loop of information processing during memory consolidation. *Nature Neuroscience*, *20*(2), 251–259. https://doi.org/10.1038/nn.4457

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology*, *13*(9), e1005768. https://doi.org/10.1371/JOURNAL.PCBI.1005768

Saunders, D. J., Patel, D., Hazan, H., Siegelmann, H. T., & Kozma, R. (2019). Locally connected spiking neural networks for unsupervised feature learning. *Neural Networks*, *119*, 332–340. https://doi.org/10.1016/j.neunet.2019.08.016

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. https://doi.org/10.1007/978-981-13-7403-6_11

Shi, C., Lu, J., Wang, Y., Li, P., & Tian, M. (2021). Exploiting Memristors for Neuromorphic Reinforcement Learning. *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 1–4. https://doi.org/10.1109/AICAS51828.2021.9458542

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Skelin, I., Hakstol, R., VanOyen, J., Mudiayi, D., Molina, L. A., Holec, V., Hong, N. S., Euston, D. R., McDonald, R. J., & Gruber, A. J. (2014). Lesions of dorsal striatum eliminate lose-switch responding but not mixed-response strategies in rats. *European Journal of Neuroscience*, *39*(10), 1655–1663. https://doi.org/10.1111/ejn.12518

Sun, C., Chen, Q., Chen, K., He, G., Fu, Y., & Li, L. (2022). Unsupervised Learning Based on Temporal Coding Using STDP in Spiking Neural Networks. *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2142–2146. https://doi.org/10.1109/ISCAS48785.2022.9937812

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *bioRxiv*, 2021.10.13.463791. https://doi.org/10.1101/2021.10.13.463791

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An Introduction (2nd edition 2018)* (Vol. 3).

Suvrathan, A. (2019). Beyond STDP — towards diverse and functionally relevant plasticity rules. *Current Opinion in Neurobiology*, *54*, 12–19. https://doi.org/10.1016/j.conb.2018.06.011

Thomas, A. (2013). Memristor-based neural networks. *Journal of Physics D: Applied Physics*, *46*(9), 093001. https://doi.org/10.1088/0022-3727/46/9/093001

Thorndike, E. L. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4). https://doi.org/10.2307/1415413

Triche, A., Maida, A. S., & Kumar, A. (2022). Exploration in neo-Hebbian reinforcement learning: Computational approaches to the exploration–exploitation balance with bio-inspired neural networks. *Neural Networks*, *151*, 16–33. https://doi.org/10.1016/j.neunet.2022.03.021

Vamplew, P., Smith, B. J., Källström, J., Ramos, G., Rădulescu, R., Roijers, D. M., Hayes, C. F., Heintz, F., Mannion, P., Libin, P. J. K., Dazeley, R., & Foale, C. (2022). Scalar reward is not enough: a response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, *36*(2), 41. https://doi.org/10.1007/s10458-022-09575-5

van der Meer, M. A., Johnson, A., Schmitzer-Torbert, N. C., & Redish, A. D. (2010). Triple Dissociation of Information Processing in Dorsal Striatum, Ventral Striatum, and Hippocampus on a Learned Spatial Decision Task. *Neuron*, *67*(1), 25–32. https://doi.org/10.1016/j.neuron.2010.06.023

van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, *11*(1), 4069. https://doi.org/10.1038/s41467-020-17866-2

van Vreeswijk, C., & Sompolinsky, H. (1997). Irregular Firing in Cortical Circuits with Inhibition/Excitation Balance. *Computational Neuroscience*, 209–213. https://doi.org/10.1007/978-1-4757-9800-5_34

Velez, R., & Clune, J. (2017). Diffusion-based neuromodulation can eliminate catastrophic forgetting in simple neural networks. *PLOS ONE*, *12*(11), e0187736. https://doi.org/10.1371/journal.pone.0187736

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4). https://doi.org/10.1007/bf00992698

Weidel, P., Duarte, R., & Morrison, A. (2021). Unsupervised Learning and Clustered Connectivity Enhance Reinforcement Learning in Spiking Neural Networks. *Frontiers in Computational Neuroscience*, *15*. https://doi.org/10.3389/fncom.2021.543872

Wilson, R. C., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00189

Wunderlich, T., Kungl, A. F., Müller, E., Hartel, A., Stradmann, Y., Aamir, S. A., Grübl, A., Heimbrecht, A., Schreiber, K., Stöckel, D., Pehle, C., Billaudelle, S., Kiene, G., Mauch, C., Schemmel, J., Meier, K., & Petrovici, M. A. (2019). Demonstrating Advantages of Neuromorphic Computation: A Pilot Study. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00260

Xing, J., Zou, X., & Krichmar, J. L. (2020). Neuromodulated Patience for Robot and Self-Driving Vehicle Navigation. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206642

Xing, J., Zou, X., Pilly, P. K., Ketz, N. A., & Krichmar, J. L. (2022). Adapting to Environment Changes Through Neuromodulation of Reinforcement Learning. https://doi.org/10.1007/978-3-031-16770-6_10

Yang, X., Song, Z., King, I., & Xu, Z. (2022). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1–20. https://doi.org/10.1109/TKDE.2022.3220219

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 3770. https://doi.org/10.1038/s41467-019-11786-6

Zamanillo, D., Sprengel, R., Hvalby, Ø., Jensen, V., Burnashev, N., Rozov, A., Kaiser, K. M. M., Köster, H. J., Borchardt, T., Worley, P., Lübke, J., Frotscher, M., Kelly, P. H., Sommer, B., Andersen, P., Seeburg, P. H., & Sakmann, B. (1999). Importance of AMPA Receptors for Hippocampal Synaptic Plasticity But Not for Spatial Learning. *Science*, *284*(5421), 1805–1811. https://doi.org/10.1126/science.284.5421.1805

Zannone, S., Brzosko, Z., Paulsen, O., & Clopath, C. (2018). Acetylcholine-modulated plasticity in reward-driven navigation: a computational study. *Scientific Reports*, *8*(1), 9486. https://doi.org/10.1038/s41598-018-27393-2

Ziegler, L., Zenke, F., Kastner, D. B., & Gerstner, W. (2015). Synaptic Consolidation: From Synapses to Behavioral Modeling. *Journal of Neuroscience*, *35*(3), 1319–1334. https://doi.org/10.1523/JNEUROSCI.3989-14.2015

Zoli, M., Torri, C., Ferrari, R., Jansson, A., Zini, I., Fuxe, K., & Agnati, L. F. (1998). The emergence of the volume transmission concept1Published on the World Wide Web on 12 January 1998.1. *Brain Research Reviews*, *26*(2-3), 136–147. https://doi.org/10.1016/S0165-0173(97)00048-9