Trends in Cognitive Sciences

CellPress

Series: Machine Behavior

Opinion

Epistemic Autonomy: Self-supervised Learning in the Mammalian Hippocampus

Diogo Santos-Pata,[1,6] Adrián F. Amil,[1,2,6] Ivan Georgiev Raikov,[3] César Rennó-Costa,[4] Anna Mura,[1] Ivan Soltesz,[3] and Paul F.M.J. Verschure [1,5,*]

Biological cognition is based on the ability to autonomously acquire knowledge, or epistemic autonomy. Such self-supervision is largely absent in artificial neural networks (ANN) because they depend on externally set learning criteria. Yet training ANN using error backpropagation has created the current revolution in artificial intelligence, raising the question of whether the epistemic autonomy displayed in biological cognition can be achieved with error backpropagation-based learning. We present evidence suggesting that the entorhinal–hippocampal complex combines epistemic autonomy with error backpropagation. Specifically, we propose that the hippocampus minimizes the error between its input and output signals through a modulatory counter-current inhibitory network. We further discuss the computational emulation of this principle and analyze it in the context of autonomous cognitive systems.

Highlights

Biological cognition is based on self-generated learning objectives. However, the mechanism by which this epistemic autonomy is realized by the neuronal substrate is not understood.

Artificial neural networks based on error backpropagation lack epistemic autonomy because they are mostly trained in a supervised fashion. In this respect, they face the symbol grounding problem of artificial intelligence.

We propose that the entorhinal–hippocampal complex, a brain structure located in the medial temporal lobe and central to memory, combines epistemic autonomy with intrinsically generated error gradients akin to error backpropagation.

We present evidence supporting the hypothesis that the counter-current inhibitory projections of the entorhinal–hippocampal complex implement a continuous self-supervised error minimization between network input and output.

The Problem of Epistemic Autonomy in Brains and Machines

The current revolution in artificial intelligence (AI) is driven by a relatively small set of core ideas stemming from the field of artificial neural networks (ANN), most notably, the use of learning rules like **error backpropagation** (see Glossary) that implement **gradient descent** in, so called, **deep learning (DL)** networks [1,2]. If advanced AI systems that emulate neural processing can be realized based on a few generic computational principles, the question becomes whether similar principles govern the organization of their biological counterparts.

Within the field of machine learning and computational neuroscience, this question has been addressed by considering the properties of cortical networks [3,4]. For instance, DL networks that capture core physiological features of the primate ventral visual stream can be generated by combining gradient descent and decorrelation [5]. Alternatively, the complex position-invariant response fields of the grid cells of the rodent entorhinal cortex (EC) can be acquired through supervised learning [6]. These two examples, however, also illustrate opposite approaches towards understanding neural processing. Whereas the former model is generative by creating a physiologically plausible ventral visual stream of a behaving agent by optimizing the intrinsic learning objective of variance minimization, the latter is descriptive by deriving an error measure from an *a priori* defined response target (i.e., recorded grid cell responses).

Despite their ability to generate activity patterns like the physiological responses of the brain either through learning plausible filter properties or mimicking prespecified output patterns, both approaches face challenges with respect to their biological plausibility. Indeed, the biological plausibility of DL methods has been debated for the last 30 years [7] and, for now, no clear solution seems to be in sight with respect to the mapping of the underlying algorithms to the biological substrate [8].

[1]Laboratory of Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS), Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Department of Neurosurgery, Stanford University, Stanford, CA, USA
[4]Digital Metropolis Institute, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil
[5]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

However, this debate might obscure the possibly more critical challenge of achieving **epistemic autonomy**, or the ability of biological cognitive systems to acquire knowledge in the absence of external supervision through explicit learning criteria targeting the representational substrate [9,10] (Box 1). Hence, despite spectacular progress in the third wave of ANN-driven AI and their success in superseding human performance in several tasks, they are still critically dependent on human engineering and supervision. This limitation restricts both the scalability of ANN and their relevance for our understanding of biological cognition. Here, we propose that recent insights in the learning networks of the medial temporal lobe, in particular the EC and hippocampus, shed a new light on these two challenges.

Towards a Framework of Self-Supervised Learning in the Hippocampus

In the field of machine learning a range of learning methods are used, which are classified as either supervised or unsupervised. The former methods classify input data based on known input–output relationships driven by an explicit error signal, while the latter learn to represent their inputs without such *a priori* criteria. The brain is considered an unsupervised learning system [11] because of its ability to discriminate and categorize novel stimuli and patterns without direct external supervision. However, the underlying neurophysiological mechanisms of learning appear to follow a more complex and multiscale organization, which is not easily captured in the supervised–unsupervised juxtaposition. Local unsupervised learning dynamics at the neuronal level are critically dependent on dedicated brain systems to modulate and gate plasticity at various stages of processing dependent on various error signals [12–14]. Hence, the question is how this multiscale cellular, circuit, and system level organization of learning supports epistemic autonomy and whether it implements error backpropagation to achieve it?

Although there is little understanding of how the complete bootstrapped learning architecture underlying the epistemic autonomy of the brain is organized, recent anatomical and physiological results on the organization of the **entorhinal–hippocampal complex (EHC)** suggest how it can be achieved using gradient descent-based self-supervised learning (Box 2). The EHC is believed to implement a prediction error-driven learning circuit, where the EC compares the difference between neocortical states, or the primary input, and signals produced by the hippocampal loop itself, or the reconstructed input [15] (Figure 1). Anatomically, the hippocampal circuit forms a nested loop (Figure 1A). The feed-forward information flow of the hippocampal trisynaptic pathway is mainly excitatory and comprises layer-specific projections from EC to the dentate gyrus (DG), CA3, and CA1, CA3 projections to CA1, and CA1 projections back to EC via the

Box 1. Historical Foundations of Epistemic Autonomy: The Symbol Grounding Problem in AI

In the 12th century, Adelard of Bath in his treatise *Natural Questions* laid the foundations for the emergence of empirical science by proposing that nature should be treated as an epistemologically closed system, where natural phenomena must be explained as being exclusively caused by natural agents, which later led to the notion of nature as ontologically closed. Epistemic autonomy focuses this foundational consideration on cognition and intelligence as natural phenomena, where an agent acquires knowledge without any dependence at the level of the substrate of its knowledge on other agents as, for instance, through an error gradient derived from externally defined prior learning objectives [10]. Indeed, this challenge of epistemic autonomy formed one of the main stumbling blocks of traditional symbolic AI, which dominated cognitive science from its rise in the 1950s until the late 1980s. This problem is well illustrated in Searle's classic Chinese room argument in which a sentient operator emulates a Turing machine by following a program for manipulating Chinese symbols. The generated output can lead outside observers to believe that the operator understands Chinese symbols and thus passes the Turing test of being a Chinese speaker. Yet, Searle's argument goes, the operator actually does not need to understand the symbols to perform the input–output transformation [77]. This quandary was subsequently dubbed the **symbol grounding problem** [78]. In other words, is the knowledge of an artificial system grounded in its own experience or in the prespecifications provided by its designers? We argue that if we speak of the biological plausibility of models of cognition, epistemic autonomy is a critical benchmark as well as the ability to address constraints derived from the anatomy and physiology of the brain and the behavior it generates [79].

[6]These authors contributed equally to this study

*Corresponding author.
pverschure@ibecbarcelona.eu
(P.F.M.J. Verschure).

Box 2. Achieving Epistemic Autonomy: The Hippocampal Spatial Code as a Benchmark

The dynamics of hippocampal learning have been addressed through computational modeling studies, providing insights into the transformation from grid cell tessellations to position-specific activity, as observed in place cells. Among these studies, mechanisms of neural selectivity [74,80,81], network structure [82], and rate remapping [16] have been proposed for the emergence of place cell-like rate activity and its modulation. Moreover, the role of novelty detection in driving learning within hippocampal networks has been pursued experimentally [83] and computationally [15]. Despite the diversity of models with varying degrees of self-supervision and pretuning of the network's synaptic distribution, a model that captures both pertinent features of the EHC and displays autonomous and continuous adaptation to environmental modifications has thus far not been formulated.

Building upon the theoretical and experimental insights summarized in this article we have defined such a computational model that can replicate many critical physiological benchmarks of EHC dynamics (Figure 2) [70]. Our model comprises a set of layers organized in a feed-forward architecture following the EHC trisynaptic loop and whose input signals mimic the physiological rate maps of the rodent medial and lateral EC during open field navigation (following [16]). The input and output layers represent the cortical and hippocampal signals, respectively, that converge in the EC comparator. In turn, the mismatch between both signals defines a gradient descent vector in the synaptic landscape to achieve a local minimum.

When driven by input signals mimicking those of the rodent medial and lateral entorhinal cortex layer II during spatial navigation [84,85], the model's entorhinal input–output mismatch minimization led to the development of spatially tuned cells, akin to the place cells found in the rodent [86]. Moreover, we observed that these hippocampal cells modulated their activity responding to environmental modifications consistent with the notion of hippocampal rate remapping [87]. Such environmental modifications influenced the network's population vector output. In addition, the sole manipulation of the sensory input (LEC cells) led to increases in the firing field size of reconstructed grid cells, an effect observed in rodent experiments [88], thus supporting the findings that spatial cues modulate grid cell activity [89]. Furthermore, with the sole purpose of minimizing the EC input–output mismatch error, the network was able to perform novelty detection and relearning in response to environmental modifications. For more detailed results on the model, see [70].

subiculum. This places the EC at a critical junction, on the one hand, as the pinnacle of the signal transduction pathways of the cortical sheet, and on the other, closing the EHC loop. This recurrent circuit displays a variety of spatial and temporal properties that contribute to the multiplexed coding of past, current, and future states of the animal within its environment [16], scaling to complex semantics in humans [17].

We analyze evidence that the EHC realizes a continuous and self-supervised gradient descent optimization of its learning dynamics. First, we outline how the EHC and its dynamics can be abstracted as a self-supervised process that minimizes an error as the difference between its input (EC superficial layer II to DG) and output layers (CA1 to EC deep layers V and VI). This error minimization is presumably achieved through a forward excitatory network and a **counter-current inhibitory** error-driven one. Second, we show that these properties of the EHC account for its pertinent features, including the formation of place cells, rate remapping, nonspecific grid cell field expansion, and place cell firing field elongation, as observed in the rodent. Third, we note that the self-contained EHC learning system displays intrinsic novelty detection and generalization capabilities.

Altogether, we propose that the generic computational role of the hippocampus is to continuously and autonomously minimize the input–output mismatch it detects in the signals from the neocortex via the EC through a process of gradient descent. This distinct computation performed by the counter-current architecture of the EHC can explain the main physiological properties of the hippocampus and shares key features with **autoencoder** networks [18]. However, our proposal goes beyond the simplistic interpretation of identifying the EHC as an autoencoder. Rather, we suggest that the input–output minimization carried out by an autoencoder optimized by error backpropagation captures one fundamental computational principle of the EHC: error-driven self-supervision supporting epistemic autonomy (Box 2). Importantly, it demonstrates that this core memory system is epistemically closed.

Glossary

Autoencoder: a type of artificial neural network that is normally characterized by a convergent–divergent topology and that has as a learning objective the reconstruction of its own inputs, thus minimizing an input–output mismatch error function and making it a self-supervised system.
Counter-current inhibition: inhibitory neuronal and synaptic activity that is systematically projected in a reverse direction with respect to a feed-forward excitatory pathway. These inhibitory counter-current projections seem to play an important role in modulating synaptic plasticity of their target excitatory neurons.
Deep learning (DL): a form of machine learning realized in artificial neural networks that combines multiple interconnected layers of neuron-like elements to incrementally extract higher-level features from input data through the backpropagation of the error between network generated output and a learning objective predefined by the designers of the system.
Entorhinal–hippocampal complex (EHC): the interconnected subregions within the cortico-hippocampal system forming a closed loop with forward excitatory and feedback inhibitory projections originating within the medial and lateral entorhinal cortex projecting to the hippocampus comprising the dentate gyrus, CA3, and CA1 regions.
Epistemic autonomy: the capacity of an agent to bootstrap and learn from its environment without explicit and direct external supervision at the level of its representational substrate.
Error backpropagation: an algorithm used for training multilayer artificial neural networks to optimize the weights of their connections based on a gradient defined by the difference, or error, between a predefined learning objective and the network's generated output given a specific input. The error is then iteratively backpropagated to modulate weight changes across the entire network.
Gradient descent: an iterative optimization algorithm that allows an efficient implementation of error backpropagation by changing the weights of a neural network following the gradient generated by the error relative to the learning objective.
Symbol grounding problem: a system that manipulates symbols,

Learning and Error Propagation in the EHC

In mapping the algorithms of DL networks to the brain, several possibilities have been considered, such as multicompartment integration, variations on spike time-dependent learning, and predictive coding (for reviews see [8,21]). Yet, it is not clear whether the brain implements an equivalent solution to error backpropagation and how it could realize it in a self-supervised fashion. In order to begin to resolve these issues, several questions must be answered, including the substrate of the error signal and its reference, the propagation of this error signal across the underlying network, and the specifics of the error-driven control of plasticity. In the following sections, we review critical pieces of evidence in support of our hypothesis that the EHC performs the functional equivalent of error backpropagation to support epistemic autonomy.

needs to define proper ontological references for these symbols in order to define their meaning. Symbolic artificial intelligence found this reference in the knowledge of the designers of these systems rather than in the experience of the system itself. Hence, it is deemed a problem.

Considering the Comparator Function of the EC as an Error Signal

Any degree of autonomous learning requires periodic assessment of the network's performance without relying on externally labeled data or error signals. This raises the question of what the substrate of the error signal may be within the EHC.

The signal flow of the hippocampal formation involves a closed-loop circuit with cortical signals being projected onto entorhinal neurons and sequentially transformed through the trisynaptic loop and finally returning to the EC (Figure 1A). This organization puts the EC in the position to gate and compare both the 'cortical-input' and 'hippocampal-output' signals [15]. The EC comparator performs a continuous computation of the error between cortically derived signals reflecting states of both the environment [lateral entorhinal cortex (LEC)] and the agent [medial entorhinal cortex (MEC)] and their associated hippocampal memory states. Indeed, novelty detection depends on the hippocampus [22] and the related error signals are believed to be computed by the EC via inhibitory competitive dynamics [23]. Notably, it has been recently observed that MEC sublayers play a key role in generating states of synchronized spiking activity supporting a neuronal mechanism for coincidence detection between cortical inputs and hippocampal outputs [24]. Novelty detection based on such a mismatch computation by necessity relies on previously stored information and the discrepancy between a cortical input state and its EHC memory reference will gate either the retrieval of previous or encoding of new experiences [25]. The former entails both pattern separation and completion within DG and CA, by virtue of the dense excitatory recurrence in CA3, facilitating the reactivation of spatiotemporal patterns similar to those observed during encoding [26].

We propose that the novelty detection attributed to the EHC, a process that is critical in regulating learning, cognitive control, and attention [9,27,28], can be better seen as an error function that serves the definition of gradient descent learning dynamics in the trisynaptic loop. Associative encoding in the EHC requires the reorganization of multiple synaptic connections along the forward excitatory hippocampal pathway. To this end, error signals generated by the EC comparator can modulate the process of learning through dopamine release within the hippocampus, enabling memory consolidation [27] as a 'print now' signal [29]. Besides the control of plasticity via dopamine release, we propose that error detection serves a second function more akin to error backpropagation: to define a gradient that shapes the overall learning dynamics. As in current and traditional machine learning models, learning occurs by updating the presynaptic weights across layers in order to minimize the error between a predefined target and actual network output [2]. Similarly, we suggest that synaptic weight updates processed in the forward excitatory hippocampal loop minimize the discrepancy between cortical inputs and their hippocampal reconstruction [21,30]. Indeed, earlier theoretical work has also demonstrated that this shift from correlation to reconstruction error-based learning is a critical transition to obtain stable performance in the combined perceptual and behavioral learning of real-world agents in the face

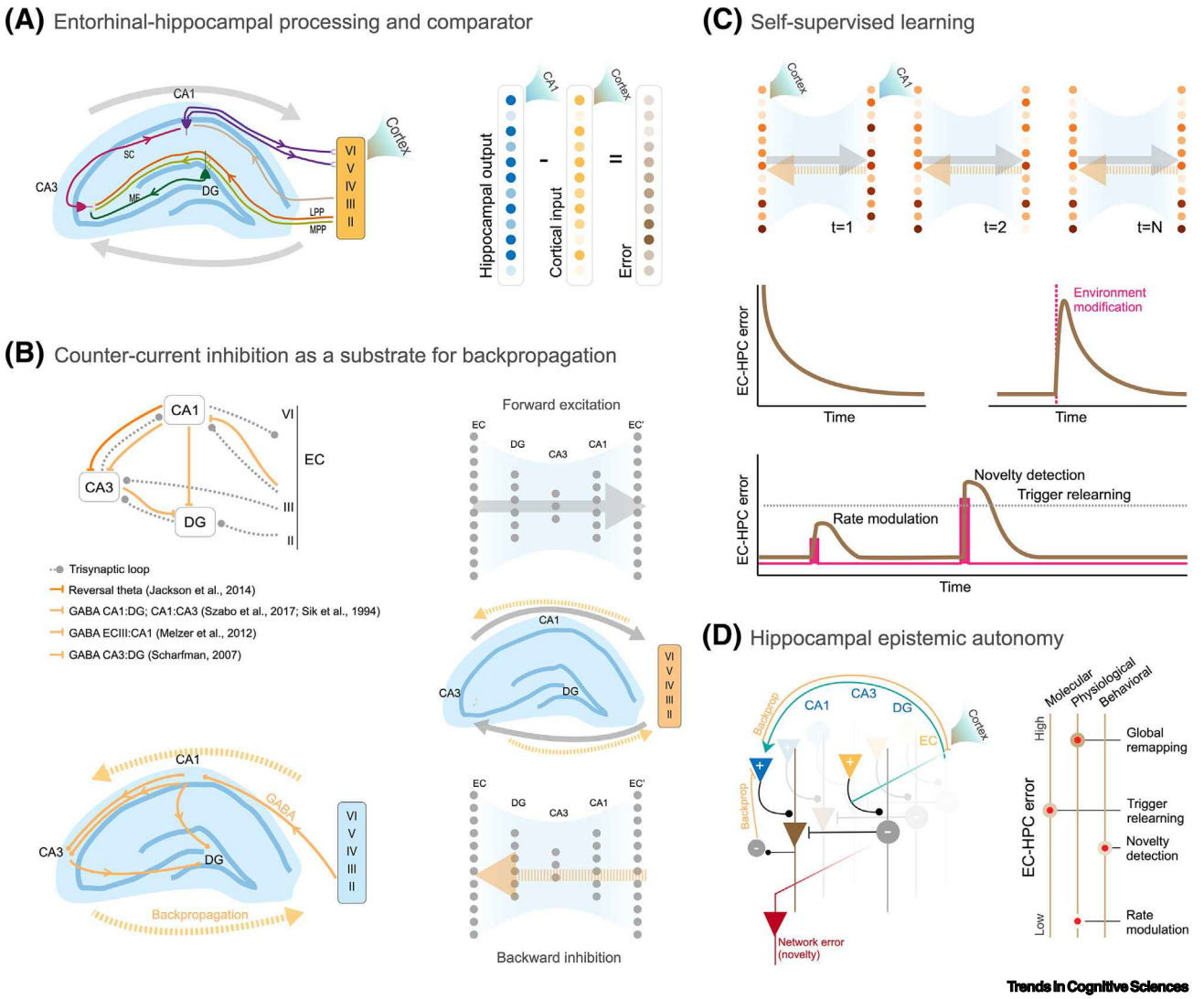

Figure 1. Forward Excitatory and Counter-Current Inhibitory Circuitry of the Entorhinal–Hippocampal Complex (EHC) Supporting Self-Supervision and Epistemic Autonomy. (A) Forward and backward hippocampal circuits. Left. The feed-forward information flow of the hippocampal trisynaptic pathway and its constituents [19]. The pathway (grey arrow) comprises projections from layer II entorhinal cortex (EC) stellate cells to the dentate gyrus (DG) and CA3 via the medial (MPP, light green) and lateral (LPP, yellow) perforant path (PP), mossy fiber projections of DG granule cells to CA3 pyramidal neurons (dark green), and CA3 projections to CA1 pyramidal neurons (the Schaffer collaterals, pink). The feedforward input is completed with direct projections from layer III EC neurons projecting to CA1 [19]. The output of the hippocampus (HPC) originates in CA1 and passes via the subiculum (not shown) to the EC LV/VI (purple). Right. Cortical input and hippocampal output coincide in EC, allowing the EHC comparator to compute the mismatch between the two signals. (B) Left. Counter-current inhibitory circuit complementing the forward excitatory loop (Box 3). Right. We hypothesize that this counter-current circuit carries error signals (yellow) that define a gradient that shapes synaptic plasticity along the forward excitatory loop, therefore implementing a biological version of error backpropagation. (C) Top. The synergy between the forward and feedback circuits shapes the continuous synaptic update in the forward loop such that it increasingly minimizes the error between the HPC input and output signals of the EC comparator. Middle. In this self-supervised learning scenario, environmental change (pink line) is reflected in the error amplitude, where error magnitude activates distinct physiological and behavioral responses. Bottom. Small amplitude errors perturbate the firing rate of principal cells, for instance, expressed as firing rate modulation in spatial navigation tasks. In contrast, large magnitude errors signal novelty and drive relearning supporting the reconstruction of this novel signal, leading to global remapping (see [20] for a possible threshold-triggered synaptic mechanism based on neuronal depolarization levels). (D) Left. The interplay between excitatory and inhibitory cells in the EHC comparator. Cortical signals coded by input neurons (yellow) are propagated throughout the trisynaptic circuit (green arrow) to neurons reflecting the HPC reconstruction of the input signal (blue). Comparator neurons (brown) receive both the reconstruction and an inhibitory copy of the input activity (grey), which in turn modulate the firing level of counter-current GABAergic interneurons that backpropagate the error signal (orange line). At this stage, recurrent GABAergic projections within the HPC modulate network-level synaptic distributions, leading to a convergence of cortical and hippocampal signals and thus performing self-supervision. Right. Dependent on its magnitude, mismatch error activates a range of molecular, physiological, and behavioral phenomena observed during spatial navigation. See [36,42,91–93].

of sampling bias and high correlations in sensory signal streams [31]. This raises the question of which substrate in the EHC could support the propagation of the error signal defined by the EC comparator. We propose that this is achieved by the counter-current inhibitory EHC network.

Counter-Current Inhibition Supports Plasticity and Error Backpropagation in the EHC

So far, we have proposed and provided evidence that learning in the EHC involves an error signal that originates in the EC comparator. However, since this error signal is the one presumably training the EHC, we ask whether there is evidence for a functional equivalent of backpropagation of error along the EHC loop and what its substrate may be?

In the standard neuroanatomical view, it has generally been held that the hippocampal formation has a canonical feed-forward information flow, as represented by the so-called hippocampal trisynaptic pathway first documented by Ramón y Cajal (Figure 1A) [32–34]. The standard view also holds that hippocampal principal cells receive feed-forward or feedback inhibition through local interneurons, with their axonal arbors restricted to the hippocampal subfield where their cell bodies reside. Principal excitatory neurons in the hippocampus by far outnumber the inhibitory interneurons [35], but the role of the latter in EHC information processing is not well understood. Here, we advance the hypothesis that EHC interneurons provide a substrate for the backpropagation of error based on their distinct anatomical and physiological properties.

Supporting this view are recent findings challenging the classical anatomical view that reveal a far more complex picture of the organization of hippocampal circuitry and the role of inhibitory interneurons in its function [36–45] (Figure 1B). Several studies indicate that some of the axons of hippocampal interneurons cross the anatomically defined boundaries of the hippocampal subfields (Box 3). These results indicate that the 'extended' interneuronal network of the hippocampal formation comprises diverse cell types beyond that of the traditionally considered interneuronal

Box 3. Substrate of Error Backpropagation in the EHC: The Counter-Current Inhibitory Network

We propose that learning in the EHC is driven by a counter-current GABAergic network. This counter-current inhibitory circuit, which complements the forward excitatory loop, consists of: (i) entorhinal layers II–III projecting back to the CA1 [91]; (ii) theta waves traveling from CA1 to CA3 [92]; (iii) feedback inhibition from the CA1 area to CA3 and DG [93]; (iv) CA3 and CA1 GABAergic neurons convey CA activity to DG [36]; and (v) deep layers of EC give rise to GABAergic projections to DG [94] (Figure 1B).

It has been shown that the axons of GABAergic neurogliaform cells in the DG not only innervate the molecular layer comprising the dendrites of the principal excitatory GCs, but also form collaterals in the adjacent CA1 and subiculum subfields [38]. A recent study using a dual retroviral and rabies virus tracing strategy showed that boundary-crossing projections constitute a mesoscale GABAergic network comprising all major GABAergic cell types [36], extending beyond local circuits and subfields. Retrograde trans-synaptic labeling of the GCs of the DG revealed nonprincipal presynaptic cells located in both the CA1 and CA3 regions (i.e., downstream in terms of the trisynaptic loop). Subsequent analysis showed that the relative numerical abundance of PV+ and SOM+ interneurons located in the CA1 and CA3 regions contributed to about 20% of the total volume of presynaptic PV/SOM GABAergic neurons innervating the GCs [36].

Further, GABAergic cells in the EC have been shown to directly project to CA1 [27,42] running counter to the standard feed-forward information flow. A similar pattern is shown by the inhibitory projections from the subiculum to CA1, reported several times over the last decades [45]. More recently, this pattern was confirmed in a study using a genetically modified rabies-based tracing method that found significant direct GABAergic inputs from the subiculum to both CA1 excitatory and inhibitory cell populations [95]. This counter-current inhibition is repeated in the interaction between CA3 and the DG. Specifically, glutamatergic excitatory pyramidal neurons in CA3 form DG projecting collaterals to hilar mossy cells and other GABAergic neurons that in turn inhibit the excitatory DG granule cells [42].

Overall, a plethora of recent anatomical and physiological studies challenge the standard view of inhibitory interneurons in the EHC operating at the local and within-subfield scale, suggesting that a complex mesoscale GABAergic network runs counter-current to the trisynaptic loop.

classes and that these neurons form a counter-current inhibitory mesoscale circuit. Therefore, these boundary-crossing projections may constitute a surprisingly robust mesoscale GABAergic system that comprises all major GABAergic cell types [36] extending beyond local within-area circuits yet more restricted than genuinely long-distance projecting cells.

The studies discussed earlier suggest that a complex boundary-crossing inhibitory counter-current network exists in the EHC that forms a potential neural substrate for error backpropagation (Figure 1B). However, a functional relationship between the counter-current inhibitory network and its effects on plasticity and learning needs to be further clarified to effectively link the former to error backpropagation. In this regard, the boundary-crossing interneurons in the CA1 and CA3 regions have been shown to relay activity back to the DG during sharp-wave ripples [36], which are high-frequency oscillatory events that are involved in episodic memory replay and have been linked to the stabilization of hippocampal place fields [46]. Other studies provide evidence of the involvement of this hippocampal inhibitory circuit in driving network-level synaptic reconfiguration and support the idea that GABAergic neurons might critically gate hippocampal plasticity. For example, the inactivation of EC somatostatin (SOM) interneurons diminishes spatial tuning of EC grid cells [47], GABA levels are correlated with memory retrieval accuracy [48], brainstem GABAergic neurons can control hippocampal contextual memories [49], and CA1 interneuron circuits are reconfigured during goal-oriented spatial learning through modification of their inputs from pyramidal cells [50] (Figure 1B). One often overlooked element of inhibitory interactions in addition to hyperpolarization is the shunting of depolarizing dendritic currents [51]. Early theoretical work had predicted that such shunting inhibition gates learning by neutralizing backpropagating action potentials required for spike time-dependent plasticity (STDP) [29]. This prediction has been experimentally confirmed in both the neocortex [52] and the hippocampus [53]. Hence, the broad counter-current inhibitory network of the hippocampus can serve as a substrate for the system level control of plasticity across all functional regions of the hippocampus through the direct control of STDP.

The mismatch between the input provided by the neocortex and the feed-forward reconstruction generated by the hippocampus has long been hypothesized to drive plasticity in the hippocampus. Further, it has been suggested that hippocampal pyramidal cells might multiplex and integrate feed-forward and feedback signals through dendritic segregation [54] (i.e., apical and basal dendrites [55]) and plateau potentials, respectively, serving the computation of the quantities driving synaptic changes as prescribed by error backpropagation [8]. Indeed, plasticity and feature selectivity in pyramidal neurons in CA1 have been shown to depend on conjunctive inputs from EC (feedback signal) and CA3 (feed-forward signal) onto different dendritic segments [56]. Furthermore, SOM+ and parvalbumin-positive (PV+) inhibitory interneurons in CA1 differentially target apical and basal dendrites of pyramidal cells, modulating feed-forward and feedback signals, respectively, while implementing distinct plasticity rules [57]. Thus, hippocampal and neocortical pyramidal neurons can detect coincident input to proximal and distal dendritic regions arising from distinct sources [58,59] and evidence suggests that neocortical signals that are integrated in the distal apical dendrites of hippocampal pyramidal cells have a strong impact on synaptic plasticity and feature selectivity [56]. Moreover, theoretical studies have also shown how these compartments can be coupled through the regulation of the dendritic length through neuromodulation [60]. In addition, hippocampal interneurons have highly convergent inputs and extensive divergent axonal projections [35], which suggests a key role in the overall gain control of the EHC [61].

These observations suggest that inhibitory interneurons may serve to minimize the error between cortical states and their hippocampal reconstruction, by dynamically gating the interactions

across dendritic compartments via plateau potentials and backpropagating action potentials [8,55] and thus modulating the spatiotemporal window in which STDP operates on coincident feed-forward and feedback signals.

The Relationship of Mismatch Error Minimization, Backpropagation, and Epistemic Autonomy

We have advanced our analysis on the premise that identifying the neuronal substrate of error backpropagation will help us understand how the brain realizes epistemic autonomy. For this, we have turned to the EHC and have shown that it performs a comparison at the level of the EC between cortical states and the hippocampal reconstruction they trigger via the forward excitatory trisynaptic loop. We have also provided evidence to support the hypothesis that the resulting error term is projected back into the hippocampal loop via a mesoscopic counter-current inhibitory stream. Finally, we have shown that plasticity in the forward excitatory loop can be modulated by core elements of this recurrent inhibitory network. Combining these elements would allow for the continuous assessment of the reconstruction error and to define a learning gradient across the EHC through inhibitory backpropagation (Figure 1C,D). As the network iterates, the mismatch error tends to decrease.

We hypothesize that error magnitude translates into the rate of changes in synaptic efficacy within the network that find an expression in molecular, physiological, and behavioral signatures of learning observed in rodent navigation (Figure 1D). In such a spatial navigation scenario, our model predicts that the EHC increasingly improves the reconstruction of the cortical input conveying environmentally relevant information such as landmarks (LEC) and their relation to movement in space (MEC). After learning, changes in the environment would lead to an increased reconstruction error in the EHC, now interpretable as novelty, which the network would progressively minimize (Figure 1C). The impact of error and novelty on the network will depend on the magnitude of the associated reconstruction error. With small environmental changes, only a slight modulation of the activity and plasticity of individual cells will occur, akin to the rate modulations in the feed-forward network observed in environmental morphing [62]. Conversely, stronger error magnitudes signal novelty, engaging the counter-current GABAergic system and inducing error-driven global remapping [63] (Figures 1C,D and 2C,F). The former case would facilitate the more accurate representation of a known environment, while the latter would initiate the acquisition of a map of a new environment. These different rate modulations triggered by novelty are well captured in our model and depend strongly on error backpropagation within the EHC (Box 2 and Figure 2). However, just because the error backpropagation algorithm turned Rosenblatt's perceptron into powerful DL models, does not necessarily mean that the brain must implement that algorithm as well. Although the core computational principles underlying hippocampal information processing remain unclear, we can point to several lines of further evidence suggesting that backpropagation-like processes provide a plausible candidate mechanism for learning in the hippocampus.

Error backpropagation is considered anatomically and biophysically implausible because it uses symmetric weights in the forward and backward directions to compute exact gradients. This contradicts neuroanatomical observations, which show a more heterogeneous organization. Yet, computational research indicates that random weights can be used in the backward direction without significantly degrading the speed or accuracy of learning [64]. Furthermore, computational studies show that error backpropagation can be approximated in ANN relying only on locally available signals and without the need to explicitly represent the error gradients [8,65]. Moreover, algorithmic analysis has shown that backpropagation is optimal on several key performance metrics of gradient descent algorithms by maximizing information rate while minimizing computational cost [66]. Thus, assuming that biological learning algorithms have

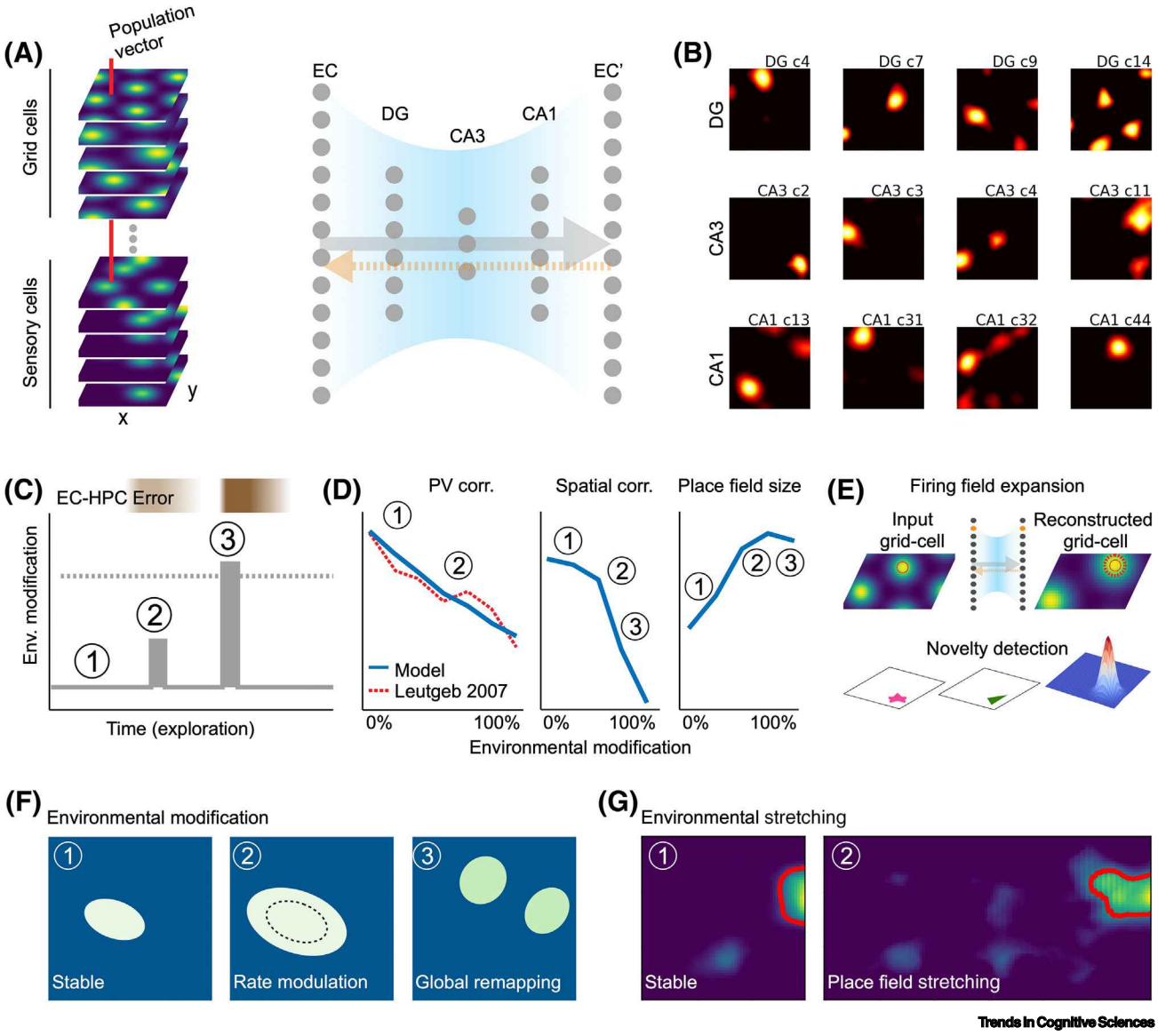

Figure 2. A Computational Model of Entorhinal–Hippocampal Complex (EHC) Self-Supervised Learning Captures Key Physiological Observations of the Rodent's Hippocampus. (A) Entorhinal rate activity as observed in the rodent during spatial navigation propagates along the hippocampal trisynaptic pathway (feed-forward flow). The match–mismatch error is computed by the difference between entorhinal cortex (EC) (cortical inputs) and EC′ (hippocampal output) following the circuit shown in Figure 1A. Backpropagation of the error via an inhibitory network serves network-level synaptic modifications to minimize the reconstruction error between EC and EC′. (B) Spatial representations and rate adaptation derived from the model's CA1 show spatially-tuned place cells similar to those found in the rodent and single cell and network-level modulation to environmental modifications reflecting rate remapping and novelty detection. (C) Effects of the magnitude of environmental change in EC reconstruction error. Navigation within a stable environment promotes learning EHC spatial representations (1) where the error magnitude determines the modulation rate of hippocampal activity (2) with a novelty threshold determining the onset of learning of a new environment (3). (D) Environmental change leads to changes in the activity of the model neurons as observed in the EHC. With increasing environmental sensory modifications, the model displays GC rate remapping equivalent [87] together with place-field expansion [88]. (E) The model replicates the increased firing field size of grid cell activity following EHC generated reconstruction [88] (Top) together with novelty detection following increased EC comparator error (Bottom). (F) Simulation of environmental changes and morphing leads to place-tuned receptive fields, rate modulation, global remapping, and field elongation. (G) The model replicates place field elongation when stretching the simulated environment along the horizontal axis [90]. See [87]. Abbreviations: DG, dentate gyrus; HPC, hippocampus; PV, parvalbumin.

optimized similar constraints of capacity and efficiency, it is reasonable to hypothesize that the brain implements approaches that are functionally similar to backpropagation in order to maximize its performance.

In addition, biologically detailed computational models inspired by the theta phase separation of encoding and retrieval [23] have shown that error-driven learning can be implemented in the EHC by allowing the EC to set a target pattern in CA1 that has to be subsequently reconstructed by the CA3 feed-forward projections at every theta cycle [67]. Indeed, this error-driven learning rule emerges from the differences in activity patterns between the peak and trough phases at each theta cycle and can be argued to approximate error backpropagation [65]. Subsequent studies have shown that such learning dynamics can lead to better performance and higher memory capacity than when the same network would only implement Hebbian plasticity [67].

Overall, given the available empirical evidence and computational results, we suggest that an approximation of error backpropagation across the trisynaptic loop instantiated by a counter-current inhibitory network, is likely to underlie self-supervised learning in the EHC.

From Hippocampal Backpropagation to Autonomous Cognitive Machines

We have outlined how the cortico-hippocampal system may combine a self-generated error signal with the gating of synaptic plasticity in the EHC following the principle of error backpropagation. The error originates in the comparison of cortical input, or scene, with its associated mnemonic reconstruction. In turn, the backpropagation of the error may be realized via an extensive counter-current inhibitory network. The most straightforward interpretation of this arrangement is captured in the 'comparator hypothesis' [15], where plasticity in the hippocampus is driven to minimize the mismatch error between its input and output. We have then argued that backpropagating the mismatch error to optimize learning leads to self-supervision that in turn explains core physiological and behavioral features of EHC (Box 2).

What does our hypothesis mean for the design of autonomous cognitive machines? Potential solutions to epistemic autonomy and its possible implementation in the brain and cognitive behaving machines must acknowledge that the system must trade-off a number of constraints, including: the consolidation of existing memories against losing them due to the formation of new memories (i.e., catastrophic forgetting [68]); its associated cost of redundancy, capacity limitations, and metabolic commitments [61]; and the recency effects or biases due to the shaping of input sampling by existing memories (i.e., behavioral feedback [31]). The error signal generated in the EC serves to satisfy these trade-offs, which are partially expressed in the phenomenon of remapping, whereby the spatial tuning of hippocampal place cells switches completely its population code when the new environment or context is significantly different from the already experienced one [69]. Indeed, our hypothesis is able to explain hippocampal remapping when a high mismatch error signal (signifying novelty) is reached, thus rapidly relearning new population codes in previously unseen environments (Box 2 and [70]). In addition, we hypothesize that the counter-current GABAergic network, besides regulating plasticity directly through inhibitory control over dendritic integration and STDP, also controls the optimization of these trade-offs through different subpopulations of interneurons. For instance, vasoactive intestinal peptide (VIP) interneurons have been shown to regulate the balance between PV+ interneurons and SOM+ interneurons, representing local and global inhibition, respectively [71]. This inter-inhibitory balance control by VIP interneurons (as well as by neuromodulators like acetylcholine) has been suggested to regulate the learning dynamics of cortical circuits [72]. Concretely, exploration and exploitation of sensory cues can be traded off on the basis of input uncertainty to find an optimal learning rate that also conserves the pre-existing sensory representations, thus providing

a potential solution to the problem of catastrophic forgetting. Hence, we propose that inter-inhibitory regulation within the counter-current GABAergic network, implementing error backpropagation in the EHC, is partly responsible for optimizing gradient descent learning according to the aforementioned efficiency and capacity trade-offs.

In addition, we can consider how the computations of the cortico-hippocampal system may overcome fundamental bottlenecks in real-time control and computation with hardware that is orders of magnitude slower than what is achieved using engineered silicon systems, while consuming only fractions of the energy budget of the hardware used for computer-based gradient descent. The brain has conserved locality in space and time in its operations, which has become a key feature in optimizing computation in the post-Moore era, where we have saturated our ability to increase the transistor density on silicon wafers [73]. The cortico-hippocampal network we describe here shows how this may be achieved, at the expense of increasing physical connectivity, in the form of dedicated inhibitory networks. An interesting optimization we can propose for machine learning algorithms based on our proposal is that the EHC embeds its slow learning dynamics in a distinct competitive fast theta-gamma code where only the most strongly feed-forward driven neurons will be active [69,74]. This implies that given an EC cortical state, only a small subset of neurons across the trisynaptic loop will be active, the synapses of which will subsequently be subject to plastic changes under the control of the recurrent inhibitory network. Hence, winner-takes-all mechanisms by specialized inhibitory interneurons enforcing decorrelation and sparseness across layers could help solve the credit assignment problem more efficiently by restricting the synaptic updates to only a selected subset of neurons. This is in stark contrast to the standard implementations of error backpropagation in ANN, which usually update all neurons and synapses at each iteration. We propose that this might partly explain the remarkable learning speeds that are displayed by the EHC in contrast to the known challenges of slow convergence in machine learning systems. In conclusion, we have outlined how one of the main targets of AI, epistemic autonomy, may be achieved in a brain system of episodic memory (Box 4), providing a plausible route to

Box 4. Error-Driven Learning throughout the Brain

Brains are learning machines that bootstrap their knowledge and behavioral policies from simple priors in interaction with the environment [96]. Although we focus on the EHC as both generating error signals and utilizing these to shape learning, other brain areas display similar processes. Indeed, it has been suggested that prediction, comparison, and error minimization are fundamental properties of information processing throughout the brain [97]. The challenge, however, is to understand the qualitative difference between the various forms of error prediction and correction that the brain displays [98]. For instance, both in the prefrontal cortex (PFC) [99] and supplementary motor area (SMA), neural dynamics are regulated by the error in the performance of goal-oriented action [100]. In addition, learning in the cerebellum is regulated by the comparison of peripheral error and its internally generated prediction [101], which can be recast as gradient descent on motor error [102]. Our analysis suggests that in all three cases epistemic autonomy is achieved through subsystem-specific error monitoring (i.e., motor error for the cerebellum, memory reconstruction error in the case of the hippocampus, and errors in achieving behavioral goals for the PFC and SMA). This raises the important question of how these different forms of error processing could be interlinked.

Notably, the EC and hippocampus share dense direct and indirect interactions with the PFC, with the PFC guiding goal-oriented recall (for a review, see [103]). We propose that the system-level epistemic autonomy of cognitive agents results from the interaction between subsystems that each are epistemically closed yet constrained by other systems. For instance, given our proposed framework, we can hypothesize that the PFC can exert top-down control over episodic memory not only biasing memory recall but also memory formation itself relative to the goals of the agent. This hypothesis is supported by the direct coupling of the PFC top-down pathway with the inhibitory counter-current circuit at the level of the EC [104]. Further, the hippocampus coordinates mnemonic responses in the cortex [105], which in turn drive the EC inputs. Hence, we see the EHC as embedded in a more elaborate system where the learning dynamics are driven by a broader interaction between the hippocampus and the cortex. From this perspective, we can speculate that the mesoscopic inhibitory counter-current network of the cortico-hippocampal system is a generic substrate for self-supervised gradient descent learning in the context of the overall brain architecture comprising dynamically coupled perceptual, cognitive, and motor systems [106] that each in turn are organized as epistemically closed.

take the hurdle of the **symbol grounding problem** and to speak of truly autonomous AI systems.

Concluding Remarks

A fundamental feature of biological cognition and a major challenge for artificial systems is epistemic autonomy. We hypothesize that the cortico-hippocampal system provides an example of how evolution might have solved this challenge. Specifically, that the EHC may function as a self-contained gradient descent learning system, continuously minimizing the input–output mismatch between states of the neocortex represented by the EC and its associated hippocampal engram. Here the engram is stored in the synapses of the excitatory forward trisynaptic loop and the error signals are propagated via a counter-current inhibitory network. Our hypothesis suggests that opposite flows of activity can thus respectively subserve recall and training in neuronal networks. This simple, yet powerful computation, which is based on a functional equivalent of error backpropagation, appears to be supported by several physiological and anatomical properties of the EHC.

Although we have presented evidence supporting the mechanisms of self-supervised learning in the hippocampus, we are still far away from having a full understanding of the EHC (see Outstanding Questions). One promising line of future research is to explore the implications of the attractor memory features of CA3 (e.g., recurrent connections performing pattern completion) combined with the pattern separation performed by DG in learning generative features of the environment, thus endowing the system with sequence generation [75] and predictive processing capabilities beyond current state of the art systems (such as variational autoencoders [76]). We have outlined how one of the fundamental problems of AI, epistemic autonomy, may be resolved within our learning framework based on prediction, comparison, and error minimization processes in which a hierarchy of distinct subsystems self-generate learning objectives while acting upon inputs that have been processed or modulated by other subsystems ultimately grounded in embodied real-world action. Hence, while in this paper we focused on the EHC, these computational principles are more generic and we predict that they are also exploited by several other brain learning systems, interaction of which shall provide a route to genuinely autonomous and biologically grounded AI systems.

Outstanding Questions

The upward and downward phases of hippocampal theta cycles (4–10 Hz) encode current and recall future locations during rodent and human spatial navigation. Does the specific theta phase set the onset of information flow of both the forward and counter-current pathways?

How is the signaling of the backpropagation of the error aligned with reversal of hippocampal theta waves, and how is the relationship of spike timing between GABAergic and pyramidal cells modulated during distinct learning stages?

What are the short- and long-term memory effects of silencing hippocampal GABAergic activity during learning and how does it vary between different interneuron subtypes?

How does the EC comparator deal with the processing delays imposed by the trisynaptic circuit and the broader cortico-hippocampal system?

How do EC and hippocampal neurons synchronize during events of memory consolidation and how strongly is this synchrony modulated by learning?

How is memory interference minimized across the distinct EHC layers and how does it evolve with learning? Moreover, how do GABAergic interneurons contribute to the separability of memory states?

Since error backpropagation relies on a graded error signal with a polarity, how does the GABAergic modulation of EHC plasticity implement the positive and negative components of such an algorithm?

Does inhibition-dependent error backpropagation provide a generic substrate to interface and regulate the core error-driven learning systems of the brain?

Acknowledgments

This article results from the 2019 Woods Hole symposium honoring the life and work of John Lisman, pioneer of system neuroscience supported by the Convergent Science Network Foundation and SPECS-lab.com. We thank the participants of the workshop for fruitful discussions and feedback on our hypothesis. In particular: György Buzsáki, Matt van der Meer, Terry Sejnowski, Marco Idiart, Ole Jensen, David Redish, and Ismael Freire. This research was supported by the European Commission Horizon 2020 grants Virtual Brain Cloud (number 826421), iNavigate (number 873178). Contributions by I.S. and I.R. were supported by an NIH BRAIN Initiative grant (U19 NS104590).

Declaration of Interests

The authors declare no conflicts of interest.

References

1. Sejnowski, T.J. (2020) The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci.* 117, 30033–30038
2. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
3. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446
4. Schmidhuber, J. (2015) Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117
5. Wyss, R. *et al.* (2006) A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.* 4, e120
6. Banino, A. *et al.* (2018) Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 429–433
7. Crick, F. (1989) The recent excitement about neural networks. *Nature* 337, 129–132
8. Lillicrap, T.P. *et al.* (2020) Backpropagation and the brain. *Nat. Rev. Neurosci.* 21, 335–346

9. Verschure, P.F.M.J. (2016) Synthetic consciousness: the distributed adaptive control perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371, 20150448
10. Verschure, P.F.M.J. (1992) *Taking connectionism seriously: the vague promise of subsymbolism and an alternative*. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society pp. 653–658
11. Dayan, P. *et al.* (1999) Unsupervised learning. In *MIT Encyclopedia of the Cognitive Sciences* (Wilson, R.A. and Keil, F., eds), MIT Press
12. Likhtik, E. and Johansen, J.P. (2019) Neuromodulation in circuits of aversive emotional learning. *Nat. Neurosci.* 22, 1586–1597
13. Ott, T. and Nieder, A. (2019) Dopamine and cognitive control in prefrontal cortex. *Trends Cogn. Sci.* 23, 213–234
14. Watabe-Uchida, M. *et al.* (2017) Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* 40, 373–394
15. Lörincz, A. and Buzsáki, G. (2000) Two-phase computational model training long-term memories in the entorhinal-hippocampal region. *Ann. N. Y. Acad. Sci.* 911, 83–111
16. Rennó-Costa, C. *et al.* (2010) The mechanism of rate remapping in the dentate gyrus. *Neuron* 68, 1051–1058
17. Solomon, E.A. *et al.* (2019) Hippocampal theta codes for distances in semantic and temporal spaces. *Proc. Natl. Acad. Sci. U. S. A.* 116, 24343–24352
18. Gluck, M.A. and Myers, C.E. (1993) Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* 3, 491–516
19. Naber, P.A. *et al.* (2001) Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum. *Hippocampus* 11, 99–104
20. Huh, D. and Sejnowski, T.J. (2018) Gradient descent for spiking neural networks. In *Advances in Neural Information Processing Systems* (Vol. 31) (Bengio, S. *et al.*, eds), pp. 1433–1443
21. Whittington, J.C.R. and Bogacz, R. (2019) Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250
22. Knight, R.T. (1996) Contribution of human hippocampal region to novelty detection. *Nature* 383, 256–259
23. Hasselmo, M.E. *et al.* (1996) Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* 6, 693–708
24. Beed, P. *et al.* (2020) Layer 3 pyramidal cells in the medial entorhinal cortex orchestrate up-down states and entrain the deep layers differentially. *Cell Rep.* 33, 108470
25. Kumaran, D. and Maguire, E.A. (2007) Match-mismatch processes underlie human hippocampal responses to associative novelty. *J. Neurosci.* 27, 8517–8524
26. O'Neill, J. *et al.* (2008) Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nat. Neurosci.* 11, 209–215
27. Duszkiewicz, A.J. *et al.* (2019) Novelty and dopaminergic modulation of memory persistence: a tale of two systems. *Trends Neurosci.* 42, 102–114
28. Rescorla, R.A. and Wagner, A.R. (1972) A theory of Pavlovian conditioning. In *Classical Conditioning II Current Research and Theory* (Black, A.H. and Prokasy, W.F., eds), pp. 497, Applteon-Century-Crofts
29. Sanchez-Montanes, M.A. *et al.* (2000) Local and global gating of synaptic plasticity. *Neural Comput.* 12, 519–529
30. Bottou, L. (1991) Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes*
31. Verschure, P.F.M.J. *et al.* (2003) Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature* 425, 620–624
32. Amaral, D.G. and Witter, M.P. (1989) The three-dimensional organization of the hippocampal formation: a review of anatomical data. *Neuroscience* 31, 571–591
33. Witter, M.P. (2006) Connections of the subiculum of the rat: topography in relation to columnar and laminar organization. *Behav. Brain Res.* 174, 251–264
34. Ramon y Cajal, S. (1911) Histologie du système nerveux de l'homme et des vertébrés. *Maloine, Paris* 2, 153–173
35. Bezaire, M.J. and Soltesz, I. (2013) Quantitative assessment of CA1 local circuits: knowledge base for interneuron-pyramidal cell connectivity. *Hippocampus* 23, 751–785
36. Szabo, G.G. *et al.* (2017) Extended interneuronal network of the dentate gyrus. *Cell Rep.* 20, 1262–1268
37. Roux, L. and Buzsáki, G. (2015) Tasks for inhibitory interneurons in intact brain circuits. *Neuropharmacology* 88, 10–23
38. Armstrong, C. *et al.* (2011) Neurogliaform cells in the molecular layer of the dentate gyrus as feed-forward γ-aminobutyric acidergic modulators of entorhinal-hippocampal interplay. *J. Comp. Neurol.* 519, 1476–1491
39. Ceranik, K. *et al.* (1997) A novel type of GABAergic interneuron connecting the input and the output regions of the hippocampus. *J. Neurosci.* 17, 5380–5394
40. Katona, L. *et al.* (2017) Behavior-dependent activity patterns of GABAergic long-range projecting neurons in the rat hippocampus. *Hippocampus* 27, 359–377
41. Lasztóczi, B. *et al.* (2011) Terminal field and firing selectivity of cholecystokinin-expressing interneurons in the hippocampal CA3 area. *J. Neurosci.* 31, 18073–18093
42. Scharfman, H.E. (2007) The CA3 "backprojection" to the dentate gyrus. *Prog. Brain Res.* 163, 627–637
43. Szabadics, J. and Soltesz, I. (2009) Functional specificity of mossy fiber innervation of GAB aergic cells in the hippocampus. *J. Neurosci.* 29, 4239–4251
44. Szabó, G.G. *et al.* (2014) Anatomically heterogeneous populations of CB_1 cannabinoid receptor-expressing interneurons in the CA3 region of the hippocampus show homogeneous input-output characteristics. *Hippocampus* 24, 1506–1523
45. Xu, X. *et al.* (2016) Noncanonical connections between the subiculum and hippocampal CA1. *J. Comp. Neurol.* 524, 3666–3673
46. Roux, L. *et al.* (2017) Sharp wave ripples during learning stabilize the hippocampal spatial map. *Nat. Neurosci.* 20, 845–853
47. Miao, C. *et al.* (2017) Parvalbumin and somatostatin interneurons control different space-coding networks in the medial entorhinal cortex. *Cell* 171, 507–521
48. Spurny, B. *et al.* (2020) Hippocampal GABA levels correlate with retrieval performance in an associative learning paradigm. *Neuroimage* 204, 116244
49. Szőnyi, A. *et al.* (2019) Brainstem nucleus incertus controls contextual memory formation. *Science* 364, eaaw0445
50. Dupret, D. *et al.* (2013) Dynamic reconfiguration of hippocampal interneuron circuits during spatial learning. *Neuron* 78, 166–180
51. Vida, I. *et al.* (2006) Shunting inhibition improves robustness of gamma oscillations in hippocampal interneuron networks by homogenizing firing rates. *Neuron* 49, 107–117
52. Letzkus, J.J. *et al.* (2011) A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* 480, 331–335
53. Müllner, F.E. *et al.* (2015) Precision of inhibition: dendritic inhibition by individual GABAergic synapses on hippocampal pyramidal cells is confined in space and time. *Neuron* 87, 576–589
54. Kaifosh, P. and Losonczy, A. (2016) Mnemonic functions for nonlinear dendritic integration in hippocampal pyramidal circuits. *Neuron* 90, 622–634
55. Guerguiev, J. *et al.* (2017) Towards deep learning with segregated dendrites. *Elife* 6, e22901
56. Bittner, K.C. *et al.* (2015) Conjunctive input processing drives feature selectivity in hippocampal CA1 neurons. *Nat. Neurosci.* 18, 1133–1142
57. Udakis, M. *et al.* (2020) Interneuron-specific plasticity at parvalbumin and somatostatin inhibitory synapses onto CA1 pyramidal neurons shapes hippocampal output. *Nat. Commun.* 11, 1–17
58. Larkum, M. (2013) A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151
59. Spruston, N. (2008) Pyramidal neurons: dendritic structure and synaptic integration. *Nat. Rev. Neurosci.* 9, 206–221
60. Verschure, P.F.M.J. and König, P. (1999) On the role of biophysical properties of cortical neurons in binding and segmentation of visual scenes. *Neural Comput.* 11, 1113–1138

61. Buzsáki, G. *et al.* (2007) Inhibition and brain work. *Neuron* 56, 771–783
62. Leutgeb, J.K. *et al.* (2005) Progressive transformation of hippocampal neuronal representations in "morphed" environments. *Neuron* 48, 345–358
63. Muller, R.U. and Kubie, J.L. (1987) The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* 7, 1951–1968
64. Lillicrap, T.P. *et al.* (2016) Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* 7, 1–10
65. O'Reilly, R.C. (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Comput.* 8, 895–938
66. Baldi, P. and Sadowski, P. (2016) A theory of local learning, the learning channel, and the optimality of backpropagation. *Neural Netw.* 83, 51–74
67. Ketz, N. *et al.* (2013) Theta coordinated error-driven learning in the hippocampus. *PLoS Comput. Biol.* 9, e1003067
68. French, R.M. (1999) Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 128–135
69. Rennó-Costa, C. *et al.* (2014) A signature of attractor dynamics in the CA3 region of the hippocampus. *PLoS Comput. Biol.* 10, e1003641
70. Santos-Pata, D. *et al.* (2021) A computational model of self-supervised learning in the hippocampus. *iScience* Published online March 26, 2021. https://doi.org/10.1016/j.isci.2021.102364
71. Fu, Y. *et al.* (2014) A cortical circuit for gain control by behavioral state. *Cell* 156, 1139–1152
72. Puigbò, J.-Y. *et al.* (2020) Switching operation modes in the neocortex via cholinergic neuromodulation. *Mol. Neurobiol.* 57, 139–149
73. Leiserson, C.E. *et al.* (2020) There's plenty of room at the top: what will drive computer performance after Moore's law? *Science* 368, eaam9744
74. De Almeida, L. *et al.* (2009) A second function of gamma frequency oscillations: an E%-max winner-take-all mechanism selects which cells fire. *J. Neurosci.* 29, 7497–7503
75. Buzsáki, G. and Tingley, D. (2018) Space and time: the hippocampus as a sequence generator. *Trends Cogn. Sci.* 22, 853–869
76. Kingma, D.P. and Welling, M. (2014) Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*
77. Searle, J.R. (1980) Minds, brains and programs. *Behav. Brain Sci.* 3, 417–424
78. Harnad, S. (1990) The symbol grounding problem. *Phys. D* 42, 335–346
79. Verschure, P.F.M.J. (2018) The architecture of mind and brain. In *Living machines: A Handbook of Research in Biomimetics and Biohybrid Systems* (Presott, T.J. *et al.*, eds), Oxford University Press
80. De Almeida, L. *et al.* (2009) The input-output transformation of the hippocampal granule cells: from grid cells to place fields. *J. Neurosci.* 29, 7504–7512
81. Savelli, F. and Knierim, J.J. (2010) Hebbian analysis of the transformation of medial entorhinal grid-cell inputs to hippocampal place fields. *J. Neurophysiol.* 103, 3167–3183
82. Cheng, S. and Frank, L.M. (2011) The structure of networks that produce the transformation from grid cells to place cells. *Neuroscience* 197, 293–306
83. Kumaran, D. and Maguire, E.A. (2007) Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus* 17, 735–748
84. Hafting, T. *et al.* (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806
85. Deshmukh, S.S. and Knierim, J.J. (2011) Representation of non-spatial and spatial information in the lateral entorhinal cortex. *Front. Behav. Neurosci.* 5, 69
86. O'Keefe, J. and Conway, D.H. (1978) Hippocampal place units in the freely moving rat: why they fire where they fire. *Exp. Brain Res.* 31, 573–590
87. Leutgeb, J.K. *et al.* (2007) Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* 315, 961–966
88. Barry, C. *et al.* (2012) Grid cell firing patterns signal environmental novelty by expansion. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17687–17692
89. Savelli, F. *et al.* (2017) Framing of grid cells within and beyond navigation boundaries. *Elife* 6, e21354
90. O'Keefe, J. and Burgess, N. (1996) Geometric determinants of the place fields of hippocampal neurons. *Nature* 381, 425–428
91. Melzer, S. *et al.* (2012) Long-range-projecting gabaergic neurons modulate inhibition in hippocampus and entorhinal cortex. *Science* 335, 1506–1510
92. Jackson, J. *et al.* (2014) Reversal of theta rhythm flow through intact hippocampal circuits. *Nat. Neurosci.* 17, 1362–1370
93. Sik, A. *et al.* (1994) Inhibitory CA1-CA3-hilar region feedback in the hippocampus. *Science* 265, 1722–1724
94. Deller, T. *et al.* (1996) A novel entorhinal projection to the rat dentate gyrus: direct innervation of proximal dendrites and cell bodies of granule cells and GABAergic neurons. *J. Neurosci.* 16, 3322–3333
95. Sun, Y. *et al.* (2014) Cell-type-specific circuit connectivity of hippocampal CA1 revealed through cre-dependent rabies tracing. *Cell Rep.* 7, 269–280
96. Pavlov, I. (1927) *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*, Oxford University Press
97. Friston, K. (2005) A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836
98. Herreros, I. and Verschure, P.F. (2015) About the goal of a goals' goal theory. *Cogn. Neurosci.* 6, 218–219
99. Matsumoto, M. *et al.* (2007) Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.* 10, 647–656
100. Marcos, E. *et al.* (2013) Neural variability in premotor cortex is modulated by trial history and predicts behavioral performance. *Neuron* 78, 249–255
101. Maffei, G. *et al.* (2017) The perceptual shaping of anticipatory actions. *Proc. R. Soc. B Biol. Sci.* 284, 20171780
102. Herreros-Alonso, I. and Arsiwalla, X.D. (2016) A forward model at Purkinje cell synapses facilitates cerebellar anticipatory control. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (Lee, D. *et al.*, eds), pp. 3828–3836, NIPS
103. Eichenbaum, H. (2017) Prefrontal–hippocampal interactions in episodic memory. *Nat. Rev. Neurosci.* 18, 547
104. Anderson, M.C. *et al.* (2016) Prefrontal–hippocampal pathways underlying inhibitory control over memory. *Neurobiol. Learn. Mem.* 134, 145–161
105. Pacheco Estefan, D. *et al.* (2019) Coordinated representational reinstatement in the human hippocampus and lateral temporal cortex during episodic memory retrieval. *Nat. Commun.* 10, 2255
106. Verschure, P.F.M.J. *et al.* (2014) The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos. Trans. R. Soc. B Biol. Sci.* 369, 20130483