CellPress
REVIEWS

Opinion

Nonmonotonic Plasticity: How Memory Retrieval Drives Learning

Victoria J.H. Ritvo,[1] Nicholas B. Turk-Browne,[2] and Kenneth A. Norman[1,3,*]

What are the principles that govern whether neural representations move apart (differentiate) or together (integrate) as a function of learning? According to supervised learning models that are trained to predict outcomes in the world, integration should occur when two stimuli predict the same outcome. Numerous findings support this, but – paradoxically – some recent fMRI studies have found that pairing different stimuli with the same associate causes differentiation, not integration. To explain these and related findings, we argue that supervised learning needs to be supplemented with unsupervised learning that is driven by spreading activation in a U-shaped way, such that inactive memories are not modified, moderate activation of memories causes weakening (leading to differentiation), and higher activation causes strengthening (leading to integration).

Highlights

Multivariate fMRI pattern analysis methods make it possible to track how neural representations change as people learn, providing new opportunities for testing theories of learning and memory.

Some recent fMRI results fit with supervised learning theories, which predict that linking two stimuli to the same associate will make their neural representations more similar (integration). However, other fMRI studies have found that representations of stimuli become less similar when linked to the same associate (differentiation).

To explain these and related findings, we argue that supervised learning needs to be supplemented with unsupervised learning that is driven by spreading activation in a U-shaped way, such that inactive memories are not modified, moderate activation of memories causes weakening (leading to differentiation), and higher activation causes strengthening (leading to integration).

Understanding Representational Change

How do stored memories change as a function of experience? The neural representations of past events can get stronger or weaker individually, but can also change with respect to each other, with neural overlap either decreasing (differentiation) or increasing (integration). These changes to the similarity structure of memories have an enormous effect on subsequent retrieval, affecting how much memories compete (less overlap results in less competition and thus better recall of distinctive features [1]) and how much generalization occurs (more overlap leads to more generalization [2]).

Here we address two fundamental, interrelated questions: (i) What are the learning rules that govern how representations change in the brain? and (ii) according to these rules, which situations lead to differentiation versus integration? These questions have come to the forefront of learning and memory research in recent years, driven by new, multivariate functional magnetic resonance imaging (fMRI) analysis methods that make it possible to track how neural similarity structure changes with learning [3]. These new **representational similarity analysis** (see Glossary) methods [4] have led to a wealth of new fMRI results that have proven to be highly constraining about underlying learning mechanisms. Some of these new results are well explained by classic **supervised learning** models, which posit that the brain adjusts representations to predict outcomes in the world [5,6]. Consistent with these theories, such studies find integration when two stimuli predict the same outcome and differentiation when two stimuli predict different outcomes (e.g., [7–9]). However, contrary to these findings, other fMRI studies have found that linking two stimuli to the same associate leads to differentiation rather than integration [10–12].

What do these seemingly contradictory findings tell us about the underlying neural learning rules? We argue that explaining representational change requires supplementing supervised learning rules with **unsupervised learning** rules that adjust neural representations based on how strongly memories are activated during the retrieval process. Furthermore, we argue that the function relating memory activation to learning is U-shaped, such that inactive memories are

[1] Department of Psychology, Princeton University, Princeton, NJ 08540, USA
[2] Department of Psychology, Yale University, New Haven, CT 06520-8205, USA
[3] Princeton Neuroscience Institute, Princeton University, Washington Road, Princeton, NJ 08544, USA

*Correspondence:
knorman@princeton.edu (K.A. Norman).

https://doi.org/10.1016/j.tics.2019.06.007

not modified, moderately activated memories are weakened, and highly activated memories are strengthened.

Supervised Learning and Representational Change

As noted earlier, a very influential account of representational change is supervised, error-driven learning. According to this view, the goal of learning is to minimize prediction error: the discrepancy between predicted and actual outcomes in the world. This idea has been instantiated in models that apply error-driven learning algorithms (e.g., error backpropagation [5]) to multilayer ('deep') neural networks. These models have experienced a resurgence of popularity in recent years as a result of successful applications to large-scale problems in artificial intelligence (AI) (e.g., computer vision [13]) and neuroscience (e.g., modeling the ventral visual stream [14,15]). Substantial progress has also been made in considering how these algorithms could be implemented in the brain [16–21].

Supervised learning models adaptively re-represent the input patterns in hidden layers of the network (between the input and output layers) to facilitate mapping from inputs to outputs. Learning rules like backpropagation adjust connection strengths throughout the network to minimize prediction error in the output layer. These adjustments can have the effect of changing the similarity structure of hidden layer patterns evoked by inputs, pushing them together or pulling them apart. In these models, if two inputs map onto the same output, error-driven learning pushes hidden layer representations of the inputs together; if the inputs map onto different outputs, error-driven learning pulls their hidden representations apart [6] (Box 1). Importantly, these representational changes are incremental in nature. The effect of supervised learning is to adjust hidden layer representations just enough to support accurate prediction, which requires sensitivity to both similarities and differences in what the inputs predict [22].

A key feature of supervised learning is its asymmetry. In these models, learning to predict distinct outputs (e.g., two different-looking faces) from similar inputs (e.g., two similar-looking scenes) yields different representational change effects than the opposite (learning to predict similar scenes from distinct faces). As such, when testing the predictions of supervised learning models on actual experimental data, it is important to identify which part of the experimental design

Glossary

Adaptive design optimization: methods that change the design of an experiment on-line (i.e., during the experiment) to better estimate quantities of interest.

Representational similarity analysis: a form of fMRI data analysis that involves comparing the spatial patterns of brain activity evoked by different stimuli or conditions; it is typically done within a particular brain region of interest. This kind of analysis can tell us whether stimulus-evoked brain patterns in a given brain region are becoming more or less similar with learning.

Supervised learning: learning rules that adjust connections between neurons based on an externally-specified teaching signal. This can involve an actual teacher specifying the correct answer, or – more simply – learning to predict sensory observations (in effect, using the world as a teacher).

Unsupervised learning: learning rules that adjust connections between neurons without leveraging an externally-specified teaching signal.

Box 1. How Supervised Learning Leads to Differentiation and Integration

Supervised neural network learning algorithms adjust connections between units based on the difference between a prediction phase (where an input is presented to the network and activity spreads to the output layer) and an outcome phase (where the 'correct' output is presented to the network). Figure I depicts a three-layer, feedforward network (i.e., activation flows unidirectionally from input to output) that learns using backpropagation [5] to map inputs to outputs via a hidden layer.

In this example, the network has already been trained to map A to X; also, A and B start out with overlapping hidden representations (as indicated by strong connections from both A and B to the middle hidden unit). Now we want to train the network using backpropagation such that either B predicts X (i.e., B and A have the same predictive consequence) or B predicts Y (i.e., B and A have different predictive consequences). When B is presented to the network during the prediction phase, X becomes partially active on the output layer because B activates some of the same hidden units that A previously activated, which (in turn) had been previously linked to X. During the outcome phase, the correct output (either X or Y) is activated; learning is based on difference between output layer activity in the prediction and outcome phase.

In the B → X case (top right), X was activated less in the prediction phase than in the outcome phase. To remedy this, backpropagation strengthens weights from active hidden units to X and also from the B input to the active hidden units that are most strongly connected to X (and thus are in the best position to boost X's activation). The result of this learning is to increase the extent to which A and B project to the same hidden units. In the B → Y case (bottom right), X was activated too strongly in the prediction phase and Y was not strongly activated enough. To remedy this, backpropagation strengthens weights from active hidden units to Y, reduces weights from active hidden units to X, and also weakens weights from B to the active hidden units that are most strongly connected to X (and thus are most strongly 'responsible' for X being overactivated). The result of this learning is to reduce the extent to which A and B project to the same hidden units.

Importantly, these predictions about differentiation and integration are not limited to the feedforward architecture shown here – the predictions generalize to recurrent supervised learning algorithms that allow activation to spread in both directions. Several recurrent algorithms have been shown to mathematically approximate feedforward backpropagation [17,20], and thus the same learning outcomes should occur.

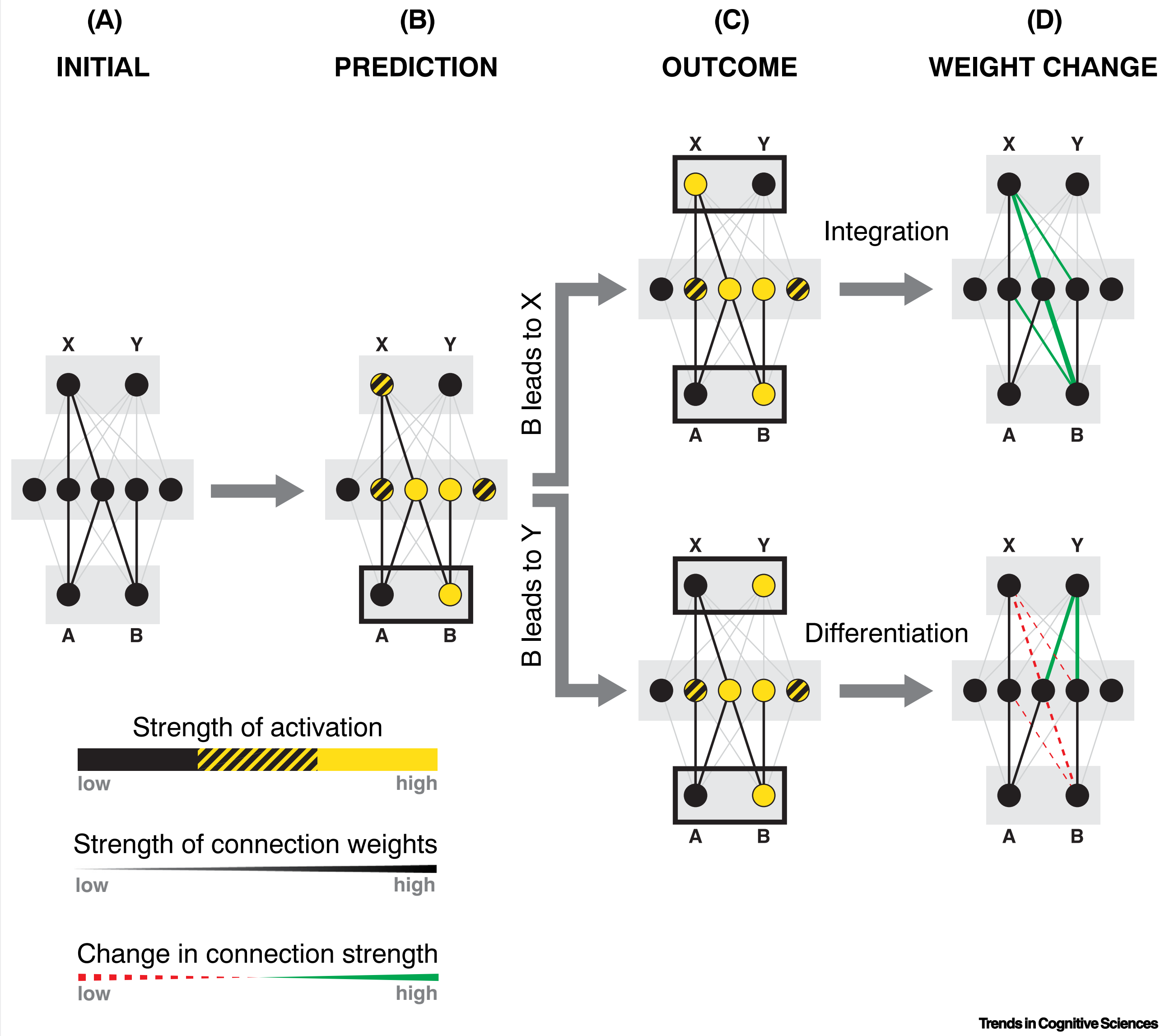


Figure I. Illustration of Supervised Learning Effects.

corresponds to the input (i.e., the cue that triggers the prediction) and which part corresponds to the output (i.e., what is being predicted). Sometimes this is straightforward (e.g., in paired-associate learning paradigms where participants are given a cue and asked to recall an associate) and sometimes less so (e.g., when to-be-associated items are presented simultaneously – in this case, participants could implicitly generate predictions by taking either one of the paired associates and trying to retrieve the other).

Numerous fMRI results support the predictions of supervised learning models about representational change. For example, one study [9] found that training participants to predict a shared scene associate from object cues (by making them imagine the object in that scene) results in integration (see also [23,24]). Supervised learning also explains how community structure affected neural representations in another study [7]. Participants viewed a sequence of arbitrary symbols, where symbols in the same community tended to transition to each other more than to symbols in other communities; neural representations of symbols within communities integrated, relative to symbols across communities, in both the cortex and the hippocampus [7,8]. Neural network models using supervised, error-driven learning (where each stimulus was used to predict the next) were able to simulate these results [7,25].

Data That Challenge Supervised Learning

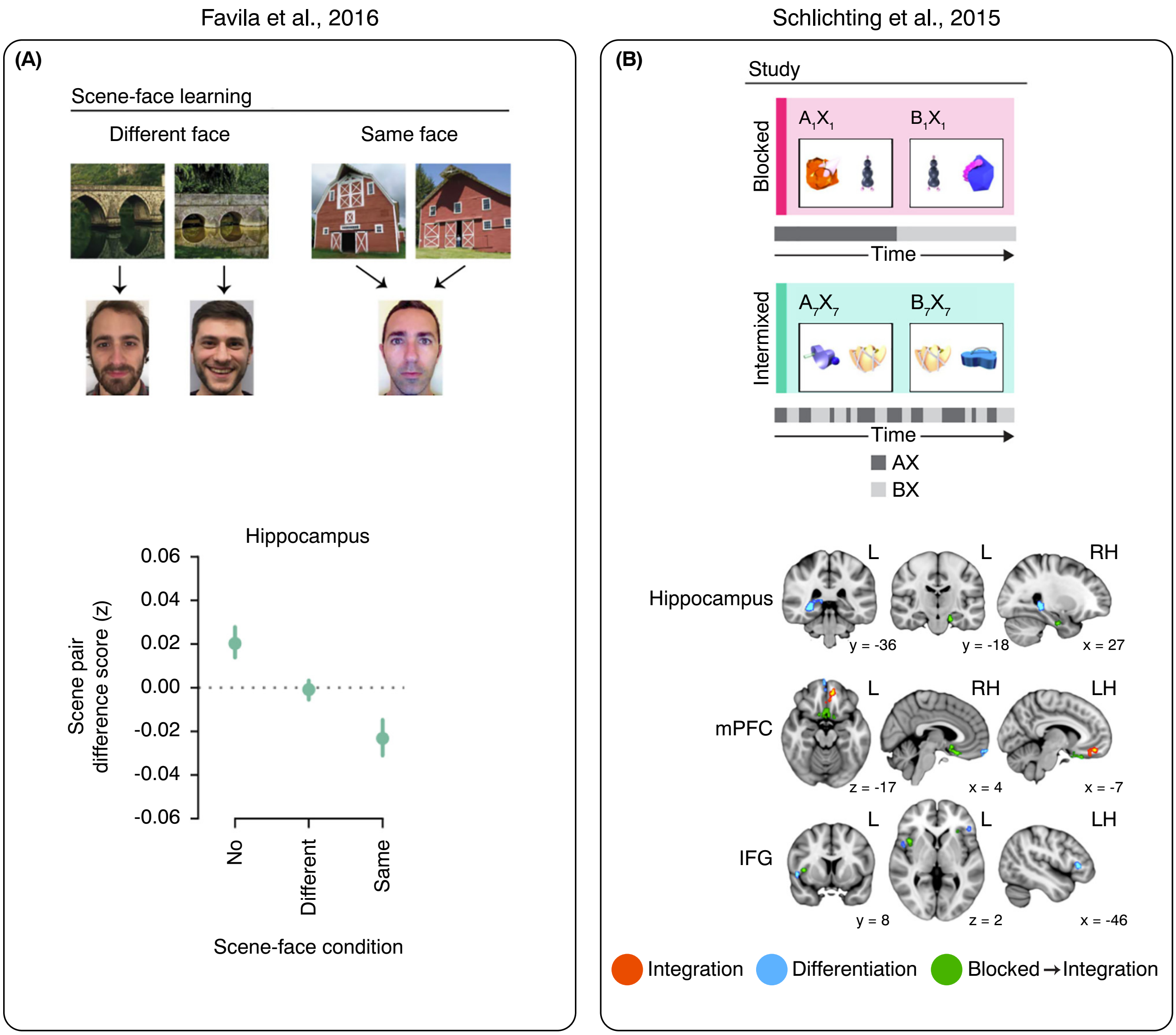
Recently, a series of fMRI studies have challenged the supervised learning account of representational change, by showing that linking memories to the same associate can sometimes lead to differentiation [10–12,26]. For example, when participants were trained to predict the same face in response to two similar scenes (e.g., two pictures of barns), this led to differentiation of the hippocampal representations of the scenes, to the point where the two scenes became less neurally similar to each other than they were to unrelated stimuli (e.g., two barns ended up as less neurally similar than a barn and a bridge) [10]. Intriguingly, this effect may vary across brain regions. One study found that linking two objects to a common associate can lead to differentiation of the objects' representations in right posterior hippocampus and integration in left medial prefrontal cortex (mPFC) [11] (Figure 1). Another study looked at the effects of presenting objects in the same episodic context (a video of a 'walkthrough' of a house) versus a different context, and found integration in one subfield of the hippocampus (CA1) but differentiation in another (CA2/3/DG) [12].

Importantly, these differentiation effects cannot simply be explained in terms of hippocampal pattern separation – the automatic bias for hippocampus to represent stimuli in a relatively orthogonalized way, driven by sparse coding [27–30]. One reason is that differentiation in the previously mentioned studies does not occur automatically, but rather only in response to the demand of learning a shared associate. Furthermore, as noted in several recent papers [10,31,32], these differentiation effects in some cases go beyond pattern separation. Orthogonalization implies that all items should have the same (low) overlap, but in [10] visually similar scenes linked to the same face had lower levels of hippocampal pattern similarity than visually dissimilar scenes (see also [31]). Lastly, whereas differentiation has often been reported in the hippocampus, it has also been observed in other regions [e.g., anterior mPFC and bilateral inferior frontal gyrus (IFG)] [11].

The aforementioned studies provide a particularly stark challenge to supervised learning models. However, it is not hard to find other memory results that pose a challenge for these supervised learning models hiding in plain sight. One example is the testing effect: the finding that successfully retrieving a memory in the absence of feedback leads to superior recall, compared with simply restudying the material [33–37]. For example, say that you learn the word pair 'absence-hollow'; later, you are given 'absence-' and you retrieve the associated word ('hollow') without feedback. From the perspective of supervised learning, one could interpret retrieval of 'hollow' as a prediction that is not matched by the world (because of the lack of feedback), so we might expect weakening of the memory based on this prediction error. Alternatively, one could argue that 'hollow' does not really constitute a prediction about what will happen in the world (since participants do not expect feedback), and hence there will be no prediction error and thus no learning. Crucially, neither of these scenarios explains why retrieval of 'hollow' is a strongly positive (reinforcing) learning experience, as has been demonstrated by myriad studies [35].

Unsupervised Learning and Representational Change

How can we explain the mix of findings described earlier? Supervised learning is too useful to discard – the vast majority of the recent successes of neural network research, in both AI and



Trends in Cognitive Sciences

Figure 1. Linking Stimuli to a Shared Associate Can Lead to Differentiation. (A) A recent fMRI study by Favila *et al.* [10] used pairs of similar scenes (e.g., two barns) as stimuli. Participants were shown scenes as cues and trained to predict faces. Specifically, participants associated each scene in a pair with a different face, the same face, or no face. Representational similarity was measured using a 'scene pair difference score': neural similarity between stimuli in the same pair (e.g., two barns), compared with stimuli in different pairs (e.g., a barn and a bridge). According to supervised learning models, stimuli with similar predictive consequences (i.e., same face scenes) should integrate [6]. However, in the hippocampus, pairmates in the same face condition showed differentiation rather than integration. Adapted from [10]. (B) In a related fMRI study by Schlichting *et al.* [11], participants encoded object pairs (e.g., AB, BC, DE, EF) such that some objects (A and C) were indirectly linked by a shared associate (B). Some related pairs were learned in a blocked fashion (all AB before any BC), whereas others were learned in an interleaved fashion. The key question was how linking to a shared associate affected the neural similarity of A and C, relative to baseline pairs that were not linked by a shared associate. The analysis focused on hippocampus, medial prefrontal cortex (mPFC), and inferior frontal gyrus (IFG). Within these regions of interest, some areas (colored blue in the figure) showed differentiation for both blocked and interleaved training (e.g., right posterior hippocampus; note that this differentiation effect was referred to as 'separation' in the original paper). Others (colored red) showed integration for both blocked and interleaved training (e.g., left mPFC). Still others (colored green) showed integration for blocked training and differentiation for interleaved training (e.g., right anterior hippocampus). Abbreviations: L, left; LH/RH, left/right hemisphere; x, y, and z coordinates, the plane of the brain slice. Adapted from [11].

neuroscience, depend on the ability to learn arbitrary input–output mappings via supervised learning [13]. Thus, rather than replacing it, we propose that supervised learning needs to be supplemented with other learning principles.

One particularly relevant set of learning principles is the class of unsupervised neural learning rules. These rules adjust synaptic connections throughout a network based on information that is local to the synapse (typically, the activation of the presynaptic and postsynaptic neurons), without any explicit consideration of how well the network is predicting external outcomes. The simplest form of unsupervised learning is the classic Hebbian rule ('fire together wire together'), which specifies that correlated firing of the presynaptic and postsynaptic neurons leads to synaptic strengthening [38,39]. This rule plays an important role in classic models of hippocampal contributions to memory (e.g., [27]). At the level of neural representations, this Hebbian rule predicts that correlated neural firing will lead to integration [40,41], even in the absence of explicit teaching signals [42–44]. There is substantial empirical support for this prediction. For example, in monkeys, increasing the correlation of inputs from the hand by surgically connecting the skin of two fingers eliminated the discontinuity between neural zones representing adjacent digits, indicating that correlated activity boosts neural similarity [45].

This simple Hebbian learning principle can also explain some findings that are challenging for supervised learning. For example, it can easily explain why retrieval of an association (e.g., retrieving 'hollow' from the cue 'absence-') reinforces that memory. With Hebbian learning, coactivity of the representations of the cue and the associate leads to strengthening of the connection between them. Importantly, however, Hebbian learning cannot straightforwardly account for the findings reviewed previously of increased differentiation for two stimuli linked to a shared associate – this should make the neurons in their representations 'fire together' more, leading to integration, not differentiation.

A more sophisticated form of unsupervised learning might do better. In particular, several learning rules have been described with the property that no learning occurs when memories are inactive, weakening occurs when memories are moderately activated, and strengthening occurs when memories are highly activated – most prominently, the Bienenstock–Cooper–Munro (BCM)

Box 2. The Nonmonotonic Plasticity Hypothesis (NMPH)

Figure IA illustrates the basic form of the nonmonotonic plasticity hypothesis (NMPH): low activation causes no learning, moderate activation causes weakening of synaptic connections, and higher levels of activation cause strengthening of synaptic connections. The most prominent computational instantiations of the NMPH (e.g., [46,87]) also incorporate metaplasticity, whereby the 'transition point' between strengthening and weakening is adjusted as a function of the average activity of the neuron; high average activity makes it easier to weaken connections into the neuron, and low average activity makes it easier to strengthen connections. This metaplasticity principle is well supported by neural data [53]; functionally, it is important for preventing runaway synaptic modification (whereby strong connections keep getting stronger).

Figure IB illustrates U-shaped plasticity in rats (adapted from [88]). Neurons in brain slices from rat visual cortex were stimulated at varying frequencies, then the slices were analyzed for long-term depression or long-term potentiation. Lower frequency stimulation produced long-term depression whereas higher frequency stimulation produced long-term potentiation.

There is also evidence of the U-shaped function in studies with humans using EEG and fMRI, showing that the NMPH scales beyond the synapse to account for human behavior. For example (Figure IC), these effects were seen in a negative priming study [50] where participants had to ignore perceptual distractors and then respond to them later. In this study, to-be-attended targets and to-be-ignored distractors came from different categories (e.g., face, scene), making it possible to track processing of the distractor using category-specific EEG pattern classifiers. Moderate levels of distractor processing were associated with negative priming (slower subsequent responding to the stimulus), but higher and lower levels of distractor processing were not associated with negative priming. Adapted from [50].

As another example (Figure ID), a study [49] obtained evidence for the NMPH in paired-associate learning, using a variant of the think/no-think paradigm [89]. Participants learned paired associates; later, they were given the cue (the first item in the pair) and asked not to think of the associated item. Moderate activation of the 'no-think' associate during the no-think trial (measured using an fMRI pattern classifier) led to diminished performance on a subsequent memory test, whereas higher activation led to enhanced performance on a subsequent memory test (adapted from [49]; the gray ribbon in the figure indicates the 90% credible interval, meaning that 90% of the probability mass of the curve distribution is contained within the interval). The relationship between memory activation and (subsequent) change in memory strength was estimated using the Probabilistic Curve Induction and Testing (P-CIT) algorithm, described in Box 4. Note that several other fMRI studies using different paradigms have obtained similar U-shaped results [54–56].

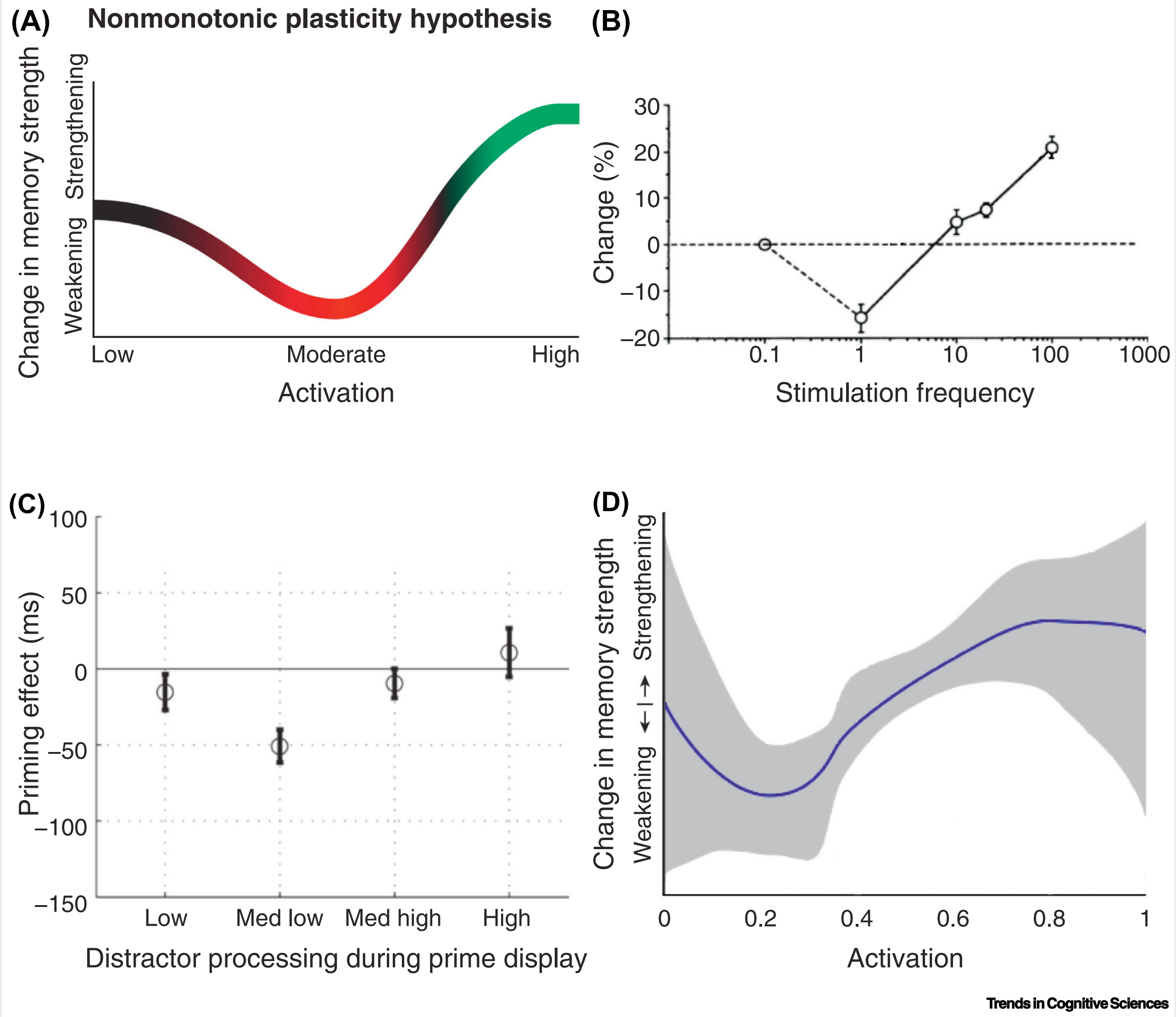


Figure I. Data Supporting the Nonmonotonic Plasticity Hypothesis.

learning rule [46,47], but also others [1,48]. We refer to the 'U-shaped' learning function shared by all of these rules as the nonmonotonic plasticity hypothesis (NMPH) [49,50]. The NMPH is supported by neurophysiological data showing that moderate postsynaptic depolarization leads to long-term depression (synaptic weakening) and stronger depolarization leads to long-term potentiation (synaptic strengthening) [51–53]. Furthermore, recent studies in humans show that the NMPH 'scales up' to explain human behavioral and neural data. Moderate activation of memories [measured using multivariate fMRI or electroencephalography (EEG)] is associated with subsequent forgetting of those memories and higher activation is associated with subsequent improvements in memory (e.g., [49,50,54–56]) (Box 2).

A useful way to think of the NMPH in relation to supervised learning is to focus on where the learning 'targets' come from. Supervised learning adapts synaptic weights so activation patterns in the network match explicit targets supplied directly by the environment. With the NMPH, the spread of activation throughout the network sets the targets. Weights are adapted to reinforce strongly activated patterns and to destabilize moderately activated patterns, at all layers of the network. In this framework, strongly activated memories (e.g., retrieving 'hollow' from the cue 'absence-') can be viewed as internally generated targets for learning.

How the NMPH Accounts for Representational Change

In addition to explaining how memories get stronger and weaker, the NMPH also makes detailed predictions about representational change [1,57]. When one memory is strongly activated and an overlapping memory is moderately activated at the same time, the NMPH predicts that connections from the strongly active memory to the moderately activated memory will be weakened, leading to differentiation; but when the overlapping memory is strongly activated, connections will be strengthened, leading to integration [57] (Figure 2, Key Figure). The functional significance of these changes is to reduce competition on subsequent retrieval attempts. This happens regardless of whether learning manifests as differentiation or integration – both reduce competition, by transforming a situation with two competing memories to either a situation with one memory (integration) or two memories that are farther apart (differentiation). Intuitively, this corresponds to two ways of preventing kids from fighting – you can make them friends (integration) or you can separate them (differentiation) [58].

This reduction in competition is also how the NMPH accounts for the testing effect. During retrieval, activation spreads to related memories, causing the representational changes outlined earlier, which reduce competition on subsequent retrieval attempts and thus improve accuracy [34]. In contrast, during restudy – which improves memory less than retrieval – activation is largely limited to the restudied memory itself. Intuitively, activation spreads less far during restudy compared with retrieval because the (highly active) restudied item laterally inhibits other memories (for a neural network model of this phenomenon see [59]). Because activation does not spread as far, there is less representational change and thus less impact on accuracy. A key prediction of this account (yet to be tested) is that the size of behavioral testing effects should correlate with the amount of representational change (e.g., measured with fMRI).

According to the NMPH, when a competing memory comes to mind moderately, it is both differentiated from the more-strongly-activated target memory and also weakened overall, making it harder to retrieve (insofar as the elements of this memory are now less-strongly interconnected and provide less mutual support to each other during retrieval). This explains the phenomenon of retrieval-induced forgetting (RIF), whereby retrieving one memory impairs the subsequent retrievability of related memories [59–62].

Furthermore, if the competing memory is subsequently restudied (or successfully retrieved, despite having been weakened), the NMPH posits that the memory will be restrengthened (due to strong coactivity of the constituent parts of the memory), while staying differentiated from the target memory. A key implication of this view is that the damaged-then-repaired memory is better off than before the damage occurred, because it has now been differentiated from the target memory and will suffer less competition from it. This leads to the prediction that memories that compete and are then restudied should be better remembered than memories that are merely restudied without having competed. This is exactly what was observed by [57,63], who termed this phenomenon 'reverse RIF'. Moreover, the size of this reversed behavioral effect can be predicted by the degree of neural differentiation of representations of the competing stimuli in the hippocampus [57]. Another prediction arising from the NMPH is that, in a RIF (or reverse RIF) paradigm, strong (as opposed to moderate) activation of a competing memory during retrieval

Key Figure

How the Nonmonotonic Plasticity Hypothesis (NMPH) Explains Integration and Differentiation

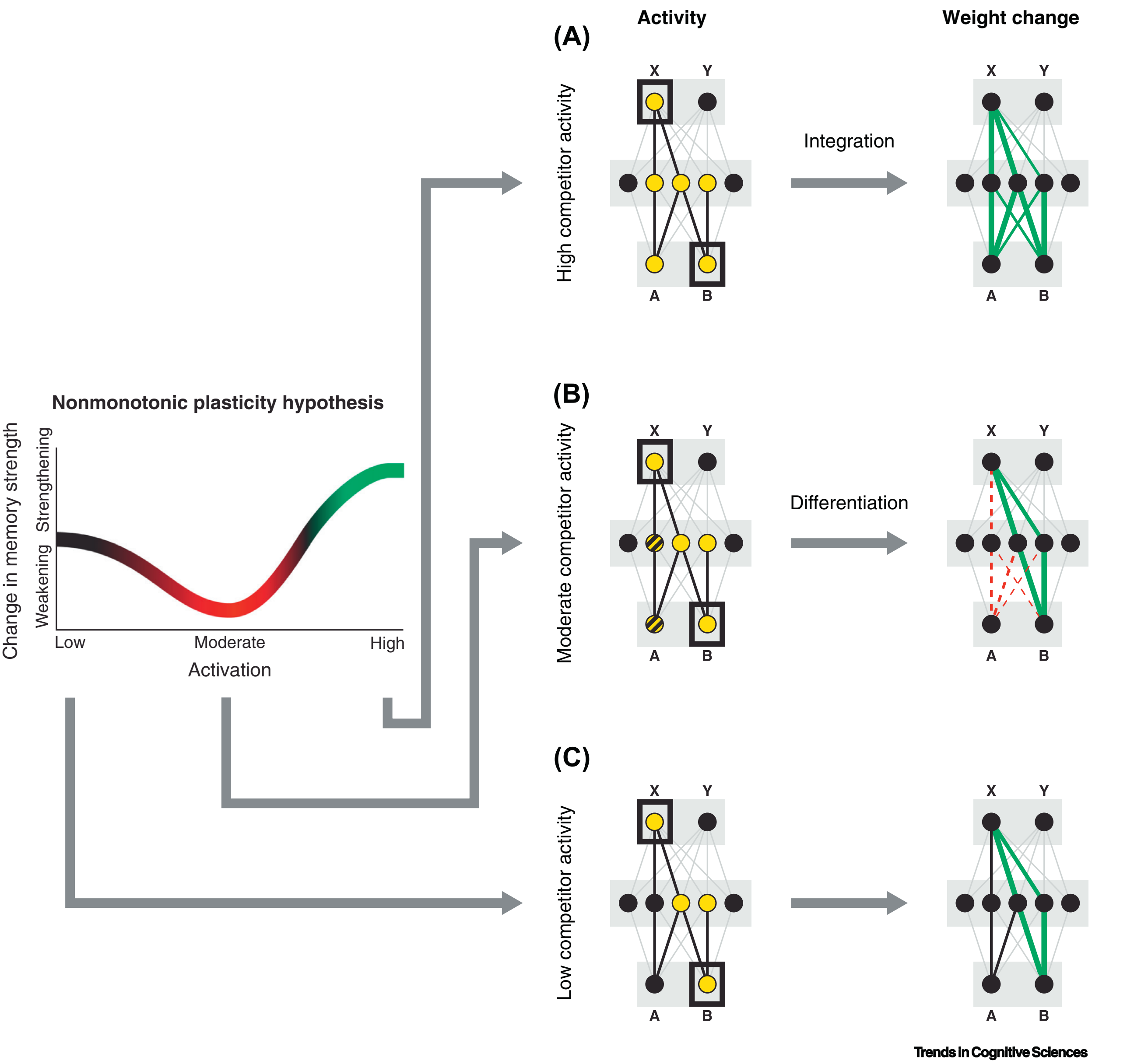


Figure 2. This figure depicts the same situation as the supervised learning box (Box 1). A has been linked to X, and B has some initial hidden layer overlap with A. In this network, activation is allowed to spread bidirectionally. When B is presented along with X (corresponding to a B–X study trial), activation can spread downward from X to the hidden layer units associated with A, and also – from there – to the input-layer representation of A. If activation spreads strongly to the input and hidden representations of A, integration of A and B occurs due to strengthening of connections between all of the strongly-activated features (top-right panel: green connections indicate strengthened weights; A–B integration can be seen by noting the increase in the number of hidden units receiving projections from both A and B). If activation spreads only moderately to the input and hidden representations of A, differentiation of A and B occurs due to weakening of connections between the moderately activated features of A and the strongly activated features of B (middle-right panel: green and red connections indicate weights that are strengthened and weakened, respectively; A–B differentiation can be seen by noting the decrease in the number of hidden units receiving strong connections from both A and B – in particular, the middle hidden unit no longer receives a strong connection from A). If activation does not spread to the features of A, then neither integration nor differentiation occurs (bottom-right panel).

should lead to integration of that memory with the retrieved memory, boosting its later accessibility. This has not been tested with fMRI, but behavioral evidence fits this prediction (e.g., [64]); see also [65] for related fMRI evidence from an associative inference paradigm.

Two recent studies of statistical learning [55,66] provide additional support for the NMPH. In these studies, participants viewed a stream of stimuli with embedded regularities (e.g., scene A was regularly followed by scene B), which were later violated (e.g., scene A was followed by face X instead of scene B). Moderate activation of the predicted scene (B) in the brain at the moment the prediction was violated led to impaired subsequent memory for B [55]. In a follow-up study [66], B was restudied after the violation; the same circumstance that led to forgetting before (moderate activation of scene B when A was unexpectedly followed by X) was associated here with neural differentiation of the hippocampal representations of A and B. This fits with the idea that moderate activation of B that co-occurs with strong activation of A leads to 'shearing' of B's connections with A (for an alternative explanation see [67]; for a related result see 'weak pairs' in [68]). Note that these statistical learning findings could also potentially be explained by supervised learning (i.e., weakening of an incorrect prediction). The place where NMPH learning diverges from supervised learning is when presenting a cue triggers very strong anticipatory activation of the predicted item and the prediction is violated. The NMPH posits that – in this case – integration of the cue and predicted item will occur, boosting the association, whereas supervised learning predicts that the association will be weakened (due to the large prediction error). In cases like this (where supervised learning and the NMPH make opposite predictions), the overall learning outcome will depend on the balance between supervised and unsupervised learning (see 'Balancing Supervised and Unsupervised Learning' section later).

Crucially, the NMPH can also explain the challenging data discussed earlier from [10,11], where linking two items to a common associate caused differentiation in some brain regions (e.g., the hippocampus); for related findings see [12,31]. In this situation, linking the items to a shared associate provides an additional conduit for activation to spread from one item to the other. If the baseline level of spreading activation was low, linking to a shared associate may push spreading activation to moderate levels, which would (according to the NMPH) lead to differentiation.

Box 3. How the NMPH Explains Findings That Challenge Supervised Learning

Figure IA shows how the NMPH can explain the hippocampal differentiation results from Favila *et al.* [10].

The network diagrams show the state of the network after scene A has been associated with face X, during a trial in which scene B is being linked either to face Y (different face condition) or face X (same face condition). In the different face condition, we assume that the level of activation spreading from scene B to its pairmate (A) in the hippocampus is relatively low. In the same face condition, the strengthened connection between X and A provides a conduit for activation to spread back down to A, resulting in A's representation becoming moderately active; this leads to differentiation of A and B, via the mechanisms shown in Figure 2 in main text.

To account for the results from Schlichting *et al.* [11], we use the same logic (Figure IB). The only extra dimension is how to account for the blocked versus interleaved training manipulation. In the blocked condition, all A–X learning occurs before any B–X trials take place; as such, the association between X and A's hidden representation will be stronger in the blocked than the interleaved condition. This strengthened A–X connection allows more activation to spread to the representation of A (in the input and hidden layers) during B–X trials.

As noted earlier, a key feature of the results from [11] is variance across brain regions as to whether integration or differentiation was observed. The NMPH can explain between-area differences in integration/differentiation in terms of between-area differences in how strongly activation spreads to A during B–X trials. The network diagrams and the topmost NMPH diagram in Figure IB depict a situation where activation spreads moderately in the interleaved condition and strongly in the blocked condition; in this situation, we would expect differentiation in the interleaved condition and integration in the blocked condition. Alternatively, if overall levels of activation are lower in a given area, one might observe differentiation in both conditions (lower NMPH diagram), and if levels of activation are higher, one might observe integration in both conditions (not shown). This account also predicts that some configurations will not be observed; in particular, if activation is higher in blocked than interleaved, there is no way to position these conditions along the NMPH curve such that blocked leads to differentiation and interleaved leads to integration. It is therefore notable that, across the three main regions of interest surveyed by [11] (hippocampus, mPFC, and IFG), all three of the 'possible' configurations were observed but the 'impossible' configuration was not observed (see Figure 1 in main text).

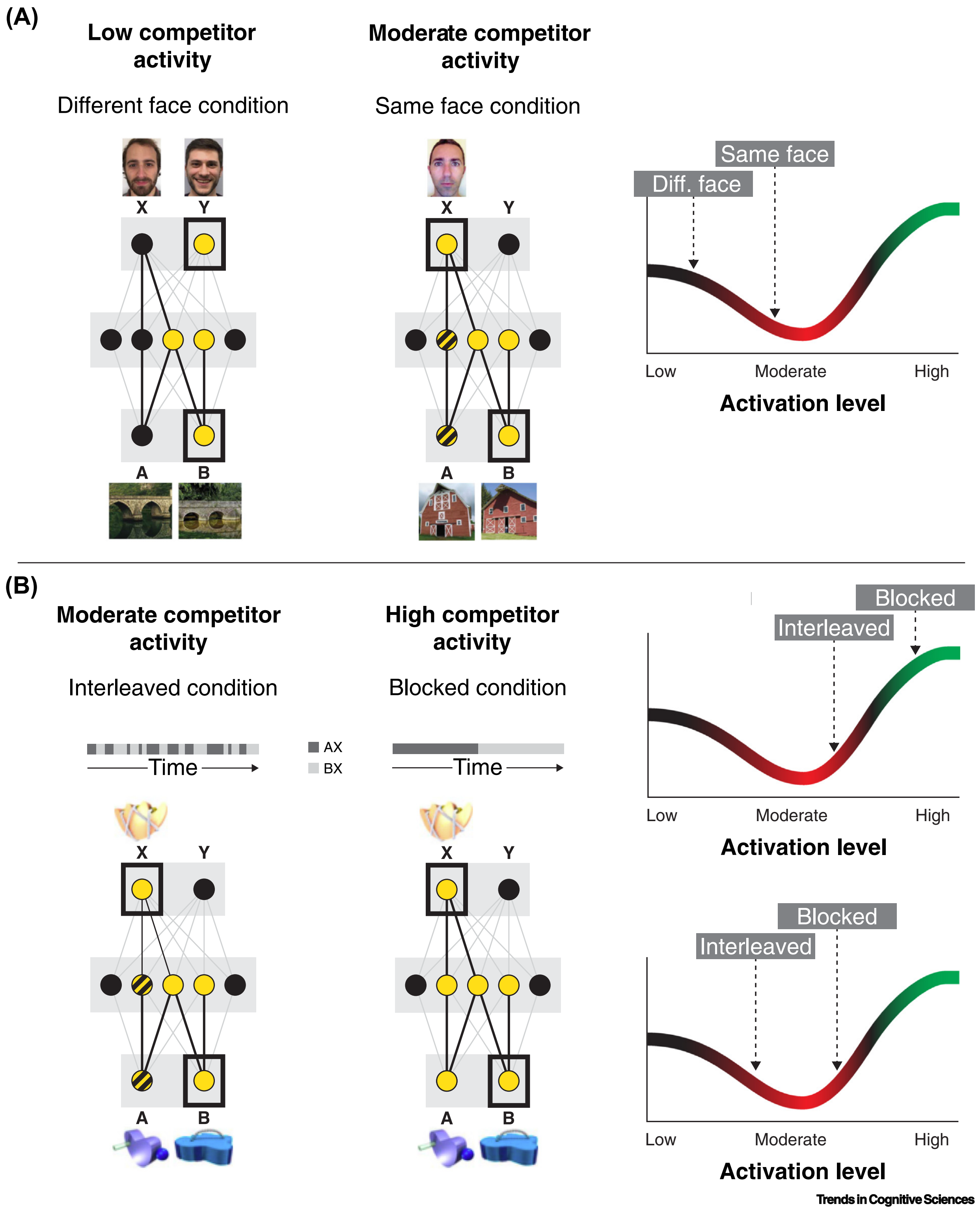


Figure I. Explanations of Challenging Findings.

However, a key implication of the NMPH is that there are limits on this dynamic; if the other item activates strongly (instead of moderately), the NMPH predicts that this will lead to integration – potentially explaining the integration effects observed in some brain regions by [11] (Box 3).

This account highlights a useful property of the NMPH: It can explain qualitative differences between brain areas (whether they tend to show integration or differentiation) in terms of a continuous, quantitative difference in how strongly competitors are allowed to come to mind. Specifically, brain areas differ in terms of how tightly excitatory activity is controlled by inhibitory interneurons (e.g., [69]). According to the NMPH, if it is relatively hard for competitors to come to mind in a particular area (because of high inhibition), that area will favor differentiation; if it is relatively easy for competitors to come to mind (because of lower inhibition), that area will favor integration. The fact that many of the differentiation effects reported in this paper involve the hippocampus may be a consequence of high inhibition and thus sparse activity in the hippocampus [28,69], which makes it difficult to strongly activate competitors. This view also fits with the aforementioned finding [12] that – within the hippocampus – subregions with very sparse activity (CA2/3/DG) [69] tend to show differentiation, whereas subregions with less sparse activity (CA1) [69] tend to show integration, as well as the finding from [11] that posterior hippocampus (hypothesized to use sparser neural codes than anterior hippocampus [70]) tends to show differentiation, whereas anterior hippocampus tends to show integration. Another factor that may contribute to differentiation in the hippocampus is that it has a high learning rate compared with cortical regions [2]. Because of this high learning rate, a single instance of competition may be enough to repel similar items a large distance apart in representational space, past the 'baseline' level of overlap for dissimilar items.

As noted by [11] (see also classic work on complementary learning systems by [2]), having different areas be differentially biased towards integration versus differentiation confers major functional benefits. If all brain regions exhibited the same level of bias towards integration or differentiation, the brain would have to choose whether to (globally) integrate or differentiate stimuli. Having brain areas with different biases towards integration versus differentiation (due to different levels of local inhibition) makes it possible to avoid this tradeoff. When multiple items are linked to a shared associate, some brain regions can integrate the items (to promote generalization) and other regions can differentiate the items (to ensure that they stay retrievable as distinct entities).

Testing the NMPH

Going forward, how do we most effectively test the NMPH? A potential criticism is that – because the NMPH is flexible enough to explain both differentiation and integration – it is not falsifiable. Consider a pair of experimental conditions (call them X and Y) where memory activation is higher in condition Y than X. If representations in condition Y show less overlap (on average) than in condition X, one could explain this *post hoc* by arguing that X and Y fall on the descending part of the U-shaped curve (where more activation leads to weakening and differentiation). Likewise, if representations in condition Y show more overlap than in condition X, one could explain this *post hoc* by arguing that X and Y lie on the ascending part of the U-shaped curve (where more activation leads to strengthening and integration). Indeed, this kind of *post hoc* theorizing can be found in Box 3. This would be less of a concern if there were an *a priori* way of knowing whether conditions would elicit low, moderate, or high activation (thereby 'locking down' whether those conditions should lead to differentiation or integration), but in practice this is hard to do.

As discussed in Box 3, the NMPH cannot explain every possible outcome; however, it still is compatible with a wide range of findings, making it challenging to test. The key to robustly testing the

NMPH account of representational change is to obtain samples from the full 'x-axis' of the U-shaped curve, thereby making it possible to reconstruct the full U-shaped pattern of plasticity predicted by the NMPH (one approach to testing for this U-shaped pattern is described in Box 4). Crucially, some results have already been obtained showing a U-shaped relationship between memory activation and subsequent memory behavior (e.g., [49]) (Box 2), providing support for the NMPH. However, a similar result has not yet been obtained in studies of representational change (showing a smooth transition from no representational change, to differentiation, to integration within a single study) – this is a goal of our ongoing research.

Combining **adaptive design optimization** methods [71] and real-time fMRI [72] may also be useful for testing the NMPH. This approach would involve tracking memory activation during fMRI scanning using real-time decoding [72,73], keeping track of which points on the activation continuum have been sampled, and adaptively changing the task to better sample regions of the continuum that have been undersampled. For example, Poppenk and Norman [74] showed that it is possible to parametrically nudge memory activation levels by varying concurrent distraction during memory retrieval using a multiple-object tracking task. This suggests an approach whereby distraction could be increased to better sample lower regions of the activation continuum if memory activation is too high, and decreased to better sample higher regions of the activation continuum if memory activation is too low. If eliminating distraction during memory retrieval is not sufficient to elicit strong activation of the memory, another option is to 'fade in' the to-be-retrieved stimulus onscreen during retrieval (for a related approach see [75]).

What is the Function of the NMPH?

Our main claim thus far has been that supplementing supervised learning with unsupervised NMPH learning may help explain data on representational change and retrieval-driven learning. Here, we revisit the question of why the brain would incorporate NMPH learning. Put another way, what functional benefits does the NMPH confer when added to supervised learning?

Box 4. Testing for a U-Shaped Curve

How do we assess whether the curve relating neural activation and plasticity has the U shape predicted by the NMPH? It helps to have multiple experimental conditions, in order to more parametrically sample the activation (x) axis. However, using multiple conditions may not be sufficient to reconstruct the full U shape; all of the conditions could, for example, fall on the descending part of the U. Also, even if there are three conditions whose mean activation values correspond to the left, middle, and right sides of the curve, it might be difficult to reconstruct the U if within-condition variability is large. For example, if the mean level of activation falls in the dip of the U, but there is variability on either side of the mean, then strengthening will occur on some trials and no learning on others, diluting the weakening effect.

Given this concern, a better approach is to exploit within-condition variability by taking all of the individual trials (which may be spread out widely across the x-axis, even if the mean levels of activation per condition are not) and then mapping neural activation values on those trials to learning outcomes. To accomplish this, we developed the P-CIT (Probabilistic Curve Induction and Testing) algorithm [49]. P-CIT takes individual trial data as input and generates (as output) a probability distribution over the space of curves relating neural activation (on the x-axis) to learning outcomes (e.g., changes in pattern similarity). P-CIT generates this probability distribution by randomly sampling curves (piecewise linear curves with three segments) and then assigning each curve an importance weight that quantifies how well the curve explains the observed relationship between neural activation and learning outcomes. Finally, these importance weights are used to compute the probability of each curve, conditioned on the observed data.

P-CIT can be used in place of simple 'binning' methods (i.e., binning trials associated with certain ranges of memory activation values, and estimating memory outcomes for these bins), which can be sensitive to the number and placement of the bin boundaries. By estimating the continuous function relating activation to memory outcomes, P-CIT avoids the need to specify bin boundaries. Another advantage of P-CIT is that it can be used to compute a Bayes factor estimate of the level of evidence for, versus against, the NMPH [54]. P-CIT curves can be computed on an individual participant basis or group basis (combining all trials across participants). Thus far, the latter approach has been used, as estimating these curves on an individual basis requires more data per participant than has been collected in past studies.

We hypothesize that the need for NMPH learning derives from the problem of competition in recurrent neural networks, where there are reciprocal interactions between layers. These networks can exhibit complex settling dynamics before reaching a stable state; if the network vacillates between nearby states without fully settling in one (as can easily happen in neural network models with noise), this can prevent the system from acting decisively in response to the current stimulus, with potentially catastrophic consequences. The NMPH could help with this problem. As noted previously, its primary functional consequence is to reduce competition on subsequent retrieval attempts, which should lead to both faster and more accurate responding (because of less vacillation, and a greater likelihood of eventually settling into the correct state).

At the same time, unsupervised learning has its costs – it can cause networks to become entrenched in particular knowledge states and make them less able to adapt to new inputs [43]; also, in pathological cases, it can lead the network to erroneous conclusions – with the NMPH, strongly retrieving a false memory can further embed that incorrect state. Our conjecture is that competition is a large enough problem to justify the need for an extra 'housekeeping' process (implemented by the NMPH) that grooms the attractor landscape to reduce competition. Specifically, we hypothesize that the potential costs of unsupervised NMPH learning (e.g., relating to entrenchment) are outweighed by these benefits conferred by that learning (in terms of reduced competition).

More simulation work is needed to assess whether this hypothesis is true. This question can be addressed by adapting recurrent neural networks that do not presently incorporate NMPH learning (e.g., the visual object recognition model described in [76]), and testing whether they benefit from the addition of the NMPH; for preliminary evidence that this is the case see [77]. These ideas build on other work describing synergies between supervised and unsupervised learning [39,78–80], including a recent model of hippocampal learning [81].

Balancing Supervised and Unsupervised Learning

One final, important question is how the brain sets the balance between supervised and unsupervised learning. As noted previously, this question is especially relevant when considering situations where the NMPH and supervised learning make conflicting predictions. Speculatively, in light of the idea that NMPH learning serves to reduce competition on future retrieval attempts, it might be useful to ramp up the influence of NMPH learning (relative to supervised learning) when competition is persistently high. Accomplishing this requires two things: a means of adjusting the influence of NMPH learning, and a control mechanism that can detect competition and trigger changes in the influence of NMPH learning.

With regard to the first issue (how to adjust the contributions of NMPH), one intriguing idea is that these contributions could be titrated by varying local levels of cortical excitability, for example, by varying the amplitude of inhibitory oscillations [1,59]. Lowering the local level of inhibition allows competing memories that are presently inactive (because inhibition outweighs excitation) to become active; this has the effect of pulling these memories from the far-left side of the U-shaped curve to the moderate activation region associated with differentiation or even the high activation region associated with integration. Thus, by increasing the extent to which competing memories come to mind, larger amplitude oscillations could lead to more representational change, which would help mitigate competition and boost subsequent memory. Although existing studies have not directly tested this claim, it is consistent with data showing that theta oscillation amplitude at encoding predicts subsequent retrieval success (e.g., [82–84]).

If the brain upregulates NMPH learning by upregulating inhibitory oscillations, what triggers that change? Under the assumption that NMPH learning should be ramped up when competition is high, it would make sense to have brain regions that are involved in detecting and resolving retrieval competition (e.g., anterior cingulate cortex and ventrolateral PFC [85]) control the strength of oscillations. Indeed, a possible path for this exists. Ventral prefrontal regions project to the basal forebrain, which controls the release of acetylcholine, which (in turn) is known to modulate oscillation strength (for a review see [86]).

Concluding Remarks and Future Directions

Determining when learning leads to differentiation versus integration has been a challenging undertaking for memory researchers. It is hard to imagine a more fundamental question, and the fMRI data reported here show that it is a complex business, full of seemingly contradictory experimental results. We have argued that supplementing supervised learning with nonmonotonic unsupervised learning helps to explain these divergent findings and a wide range of other learning phenomena, including the testing effect, (reverse) RIF, and effects of expectation–violation in statistical learning. Although there is still much more work to be done (see Outstanding Questions), the research reviewed here shows that the flow of activation through neural networks can powerfully sculpt memories that are activated, strengthening or weakening them, and pushing them together or pulling them apart, in ways that cannot always be explained by supervised learning. Now that we have the means to track both memory activation and the resulting representational change in humans using fMRI, we expect rapid progress in characterizing these unsupervised learning effects in the years to come.

Outstanding Questions

Is it possible to observe a transition from differentiation to integration, as a function of increasing competitor activation, within the same study?

What are the most sensitive ways to measure the behavioral consequences of neural differentiation and integration?

Does unsupervised, nonmonotonic learning (implemented in algorithms like BCM) benefit the overall performance of recurrent nets by reducing competition?

What determines the balance between supervised and unsupervised learning, and can it be altered?

How does this balance change over development, given rapid learning during this period, differential maturation of brain regions, and continually changing feedback from the world (as motor, language, and cognitive abilities grow)?

Does the NMPH apply equally to the whole brain, or is it especially prevalent in some structures?

What role does offline learning (e.g., during sleep or quiet wake) play in representational change, and how does the NMPH contribute to this? Unsupervised learning may be especially important (relative to supervised learning) during sleep, since the brain is cut off from the external world and sensory input cannot provide a 'target' for learning.

Can the NMPH be leveraged to develop new clinical treatments? Several clinical conditions are characterized by insufficient neural differentiation (e.g., of phonological representations, in the case of developmental dyslexics; and visual representations, in the case of patients with temporal lobe damage). A deeper understanding of representational change may provide new and better ways of redifferentiating collapsed representations and ameliorating associated cognitive deficits.

Can closed-loop neurofeedback be used to test the NMPH?

Acknowledgments

This work was supported by National Institute of Mental Health (NIMH) R01MH069456 to K.A.N. and N.B.T.B., and an National Science Foundation (NSF) Graduate Research Fellowship to V.J.H.R. We are grateful to Judy Fan, Chris Honey, Justin Hulbert, Jarrod Lewis-Peacock, Ehren Newman, Jordan Poppenk, Randy O'Reilly, Andrew Saxe, Anna Schapiro, and others for their comments on earlier drafts of this opinion article.

References

1. Norman, K.A. *et al.* (2006) How inhibitory oscillations can train neural networks and punish competitors. *Neural Comput.* 18, 1577–1610
2. McClelland, J.L. *et al.* (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. *Psychol. Rev.* 102, 419–457
3. Lewis-Peacock, J.A. and Norman, K.A. (2014) Multi-voxel pattern analysis of fMRI data. In *The Cognitive Neurosciences* (5th edn) (Gazzaniga, M.S. and Mangun, G.R., eds), pp. 911–920, MIT Press
4. Kriegeskorte, N. *et al.* (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4
5. Rumelhart, D.E. *et al.* (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Foundations, Vol. 1)* (Rumelhart, D.E. *et al.*, eds), pp. 318–362, MIT Press
6. Gluck, M.A. and Myers, C.E. (1993) Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* 3, 491–516
7. Schapiro, A.C. *et al.* (2013) Neural representations of events arise from temporal community structure. *Nat. Neurosci.* 16, 486–492
8. Schapiro, A.C. *et al.* (2016) Statistical learning of temporal community structure in the hippocampus. *Hippocampus* 26, 3–8
9. Tompary, A. and Davachi, L. (2017) Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* 96, 228–241
10. Favila, S.E. *et al.* (2016) Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat. Commun.* 7, 11066
11. Schlichting, M.L. *et al.* (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* 6, 8151
12. Dimsdale-Zucker, H.R. *et al.* (2018) CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nat. Commun.* 9, 294
13. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
14. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365
15. Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, e1003915
16. Scellier, B. and Bengio, Y. (2017) Equilibrium propagation: bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* 11, 24
17. O'Reilly, R.C. (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Comput.* 8, 895–938
18. O'Reilly, R.C. *et al.* (2017) Deep predictive learning: a comprehensive model of three visual streams. *arXiv* Published online September 14, 2017. arxiv:1709.04654
19. Keller, G.B. and Mrsic-Flogel, T.D. (2018) Predictive processing: a canonical cortical computation. *Neuron* 100, 424–435
20. Whittington, J.C.R. and Bogacz, R. (2019) Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250
21. Richards, B.A. and Lillicrap, T.P. (2019) Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* 54, 28–36

22. McClelland, J.L. and Rogers, T.T. (2003) The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* 4, 310–322
23. Milivojevic, B. *et al.* (2015) Insight reconfigures hippocampal-prefrontal memories. *Curr. Biol.* 25, 821–830
24. Collin, S.H.P. *et al.* (2015) Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* 18, 1562–1564
25. Schapiro, A.C. *et al.* (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20160049
26. Ballard, I.C. *et al.* (2019) Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* 10, 1073
27. Marr, D. (1971) Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 262, 23–81
28. O'Reilly, R.C. and McClelland, J.L. (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* 4, 661–682
29. Norman, K.A. and O'Reilly, R.C. (2003) Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol. Rev.* 110, 611–646
30. Yassa, M.A. and Stark, C.E.L. (2011) Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525
31. Chanales, A.J.H. *et al.* (2017) Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr. Biol.* 27, 2307–2317
32. Duncan, K.D. and Schlichting, M.L. (2018) Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* 153, 40–56
33. Rafidi, N.S. *et al.* (2018) Reductions in retrieval competition predict the benefit of repeated testing. *Sci. Rep.* 8, 1–12
34. Antony, J.W. *et al.* (2017) Retrieval as a fast route to memory consolidation. *Trends Cogn. Sci.* 21, 573–576
35. Rowland, C.A. (2014) The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140, 1432–1463
36. van den Broek, G. *et al.* (2016) Neurocognitive mechanisms of the 'testing effect': a review. *Trends Neurosci. Educ.* 5, 52–66
37. Karpicke, J.D. and Roediger, H.L. (2008) The critical importance of retrieval for learning. *Science* 319, 966–968
38. Grossberg, S. (1976) Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 23, 121–134
39. O'Reilly, R.C., Munakata, Y., eds (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press
40. Buonomano, D.V. and Merzenich, M.M. (1998) Cortical plasticity: from synapses to maps. *Annu. Rev. Neurosci.* 21, 149–186
41. McClelland, J.L. *et al.* (2002) Teaching the /r/–/l/ discrimination to Japanese adults: behavioral and neural aspects. *Physiol. Behav.* 77, 657–662
42. Hall, G. (1996) Learning about associatively activated stimulus representations: implications for acquired equivalence and perceptual learning. *Anim. Learn. Behav.* 24, 233–255
43. McCandliss, B.D. *et al.* (2002) Success and failure in teaching the [r]–[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cogn. Affect. Behav. Neurosci.* 2, 89–108
44. Symonds, M. and Hall, G. (1995) Perceptual learning in flavor aversion conditioning: roles of stimulus comparison and latent inhibition of common stimulus elements. *Learn. Motiv.* 26, 203–219
45. Clark, S.A. *et al.* (1988) Receptive fields in the body-surface map in adult cortex defined by temporally correlated inputs. *Nature* 332, 444–445
46. Bienenstock, E.L. *et al.* (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* 2, 32–48
47. Cooper, L.N. *et al.* (2004) *Theory of Cortical Plasticity*, World Scientific
48. Diederich, S. and Opper, M. (1987) Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.* 58, 949–952
49. Detre, G.J. *et al.* (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* 51, 2371–2388
50. Newman, E.L. and Norman, K.A. (2010) Moderate excitation leads to weakening of perceptual representations. *Cereb. Cortex* 20, 2760–2770
51. Artola, A. *et al.* (1990) Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature* 347, 69–72
52. Hansel, C. *et al.* (1996) Different threshold levels of postsynaptic [Ca^{2+}]$_i$ have to be reached to induce LTP and LTD in neocortical pyramidal cells. *J. Physiol. Paris* 90, 317–319
53. Bear, M.F. (2003) Bidirectional synaptic plasticity: from theory to reality. *Philos. Trans. R. Soc. B Biol. Sci.* 358, 649–655
54. Lewis-Peacock, J.A. and Norman, K.A. (2014) Competition between items in working memory leads to forgetting. *Nat. Commun.* 5, 5768
55. Kim, G. *et al.* (2014) Pruning of memories by context-based prediction error. *Proc. Natl. Acad. Sci.* 111, 8997–9002
56. Wang, T.H. *et al.* (2019) More is less: increased processing of unwanted memories facilitates forgetting. *J. Neurosci.* 39, 3551–3560
57. Hulbert, J.C. and Norman, K.A. (2015) Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cereb. Cortex* 25, 3994–4008
58. Paller, K.A. *et al.* (2019) Replay-based consolidation governs enduring memory storage. In *The Cognitive Neurosciences* (6th edn) (Gazzaniga, M.S. *et al.*, eds), MIT Press
59. Norman, K.A. *et al.* (2007) A neural network model of retrieval-induced forgetting. *Psychol. Rev.* 114, 887–953
60. Anderson, M.C. *et al.* (1994) Remembering can cause forgetting: retrieval dynamics in long-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 1063–1087
61. Wimber, M. *et al.* (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nat. Neurosci.* 18, 582–589
62. Murayama, K. *et al.* (2014) Forgetting as a consequence of retrieval: a meta-analytic review of retrieval-induced forgetting. *Psychol. Bull.* 140, 1383–1409
63. Storm, B.C. *et al.* (2008) Accelerated relearning after retrieval-induced forgetting: the benefit of being forgotten. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 230–236
64. Chan, J.C.K. (2009) When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *J. Mem. Lang.* 61, 153–170
65. Zeithamova, D. *et al.* (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* 75, 168–179
66. Kim, G. *et al.* (2017) Neural differentiation of incorrectly predicted memories. *J. Neurosci.* 37, 2022–2031
67. Greve, A. *et al.* (2018) Neural differentiation of incorrectly predicted memories. *Front. Hum. Neurosci.* 12, 278
68. Schapiro, A.C. *et al.* (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr. Biol.* 22, 1622–1627
69. Barnes, C.A. *et al.* (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.* 83, 287–300
70. Keinath, A.T. *et al.* (2014) Precise spatial coding is preserved along the longitudinal hippocampal axis. *Hippocampus* 24, 1533–1548
71. Myung, J.I. *et al.* (2013) A tutorial on adaptive design optimization. *J. Math. Psychol.* 57, 53–67
72. Sitaram, R. *et al.* (2017) Closed-loop brain training: the science of neurofeedback. *Nat. Rev. Neurosci.* 18, 86–100
73. deBettencourt, M.T. *et al.* (2015) Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* 18, 470–475
74. Poppenk, J. and Norman, K.A. (2017) Multiple-object tracking as a tool for parametrically modulating memory reactivation. *J. Cogn. Neurosci.* 29, 1339–1354
75. deBettencourt, M.T. *et al.* (2019) Neurofeedback helps to reveal a relationship between context reinstatement and memory retrieval. *NeuroImage* 200, 292–301

76. Tang, H. *et al.* (2018) Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. U. S. A.* 115, 8835–8840
77. O'Reilly, R.C. *et al.* (2013) Recurrent processing during object recognition. *Front. Psychol.* 4, 1–14
78. O'Reilly, R.C. *et al.* (2015) The Leabra cognitive architecture: how to play 20 principles with nature and win! In *The Oxford Handbook of Cognitive Science*In (Vol. 1) (Chipman, S.E.F., ed.), pp. 91–116, Oxford University Press
79. Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *Science* 313, 504–507
80. Love, B.C. *et al.* (2004) SUSTAIN: a network model of category learning. *Psychol. Rev.* 111, 309–332
81. Mack, M.L. *et al.* (2016) Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proc. Natl. Acad. Sci.* 113, 13203–13208
82. Klimesch, W. (1999) EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195
83. Osipova, D. *et al.* (2006) Theta and gamma oscillations predict encoding and retrieval of declarative memory. *J. Neurosci.* 26, 7523–7531
84. Sederberg, P.B. *et al.* (2003) Theta and gamma oscillations during encoding predict subsequent recall. *J. Neurosci.* 23, 10809–10814
85. Kuhl, B.A. *et al.* (2007) Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat. Neurosci.* 10, 908–914
86. Newman, E.L. *et al.* (2012) Cholinergic modulation of cognitive processing: insights drawn from computational models. *Front. Behav. Neurosci.* 6, 24
87. O'Reilly, R.C. *et al.* (2012) *Computational Cognitive Neuroscience* (1st edn), Wiki Book
88. Kirkwood, A. *et al.* (1996) *Experience-dependent modification of synaptic plasticity in visual cortex*. 381 pp. 526–528
89. Anderson, M. and Green, C. (2001) Suppressing unwanted memories by executive control. *Nature* 410, 131–134