Trends in
Cognitive Sciences

CellPress

Opinion

Demystifying unsupervised learning: how it helps and hurts

Franziska Bröker [1,2,3,4,*], Lori L. Holt[5], Brett D. Roads[6], Peter Dayan[1,7,8], and Bradley C. Love[6,8]

Humans and machines rarely have access to explicit external feedback or supervision, yet manage to learn. Most modern machine learning systems succeed because they benefit from unsupervised data. Humans are also expected to benefit and yet, mysteriously, empirical results are mixed. Does unsupervised learning help humans or not? Here, we argue that the mixed results are not conflicting answers to this question, but reflect that humans self-reinforce their predictions in the absence of supervision, which can help or hurt depending on whether predictions and task align. We use this framework to synthesize empirical results across various domains to clarify when unsupervised learning will help or hurt. This provides new insights into the fundamentals of learning with implications for instruction and lifelong learning.

Highlights

Humans are not guaranteed to benefit from unsupervised experiences (and neither are machines).

Instead, given unsupervised experience, humans self-reinforce their predictions. This can help performance when the predictions are accurate; it can hurt or have no effect when the predictions are inaccurate.

Predictions depend on the internal representations of learners, which are shaped by prior experiences. Thus, prediction accuracy depends on how well internal representations align with the task. Only by assessing these representations can researchers understand whether and why unsupervised learning helps or hurts in a specific task and in a specific person.

The literatures on self-reinforcement and unsupervised learning in humans have largely operated in isolation, but would benefit from more crosstalk.

Insights also have broad implications for lifelong learning and the design of instruction.

> “There was, Carter thought, a downside to experience. ‘Experience is making the same mistake over and over again, only with greater confidence,’ he said. The line wasn’t his, but he liked it.”
>
> [Michael Lewis, *The Premonition: A Pandemic Story*]

Supervised and unsupervised learning

We live and learn in an environment that rarely provides us with **supervision** (see Glossary) in the form of explicit external **feedback**. For example, we have learned to call some animals ‘sheep’ and others ‘goats’. Many of us acquired this distinction at a young age when we spent much time around our caretakers. Like an external teacher, they provided us explicitly with the correct labels by naming animals in our field of view. Getting older, we still encounter sheep and goats, as well as animals we have never seen before, but we now rarely have a teacher in tow. Thus, our learning about the world could be helped if we also made use of the information contained in all these unsupervised experiences (Figure 1).

Machine learning faces a conspicuously similar problem. Typically, an abundance of unsupervised data is available for learning (e.g., images of sheep and goats), but supervision (e.g., human-annotated sheep/goat labels for each image) is rare and expensive. This has led to extensive research aiming to harness the information contained in unsupervised data. As a result, we now have powerful **learning algorithms** able to extract statistical information and features from unsupervised data [1], which can be further fine-tuned to specific tasks [2] or used to boost **supervised learning** [3]. Ultimately, the tremendous success of machine learning methods stems from their ability to learn in the absence of supervision.

The mystery of unsupervised learning in humans

It appears clear that both humans and machines benefit from leveraging unsupervised experiences. Thus, there has been a surge in empirical and computational work over the past decades

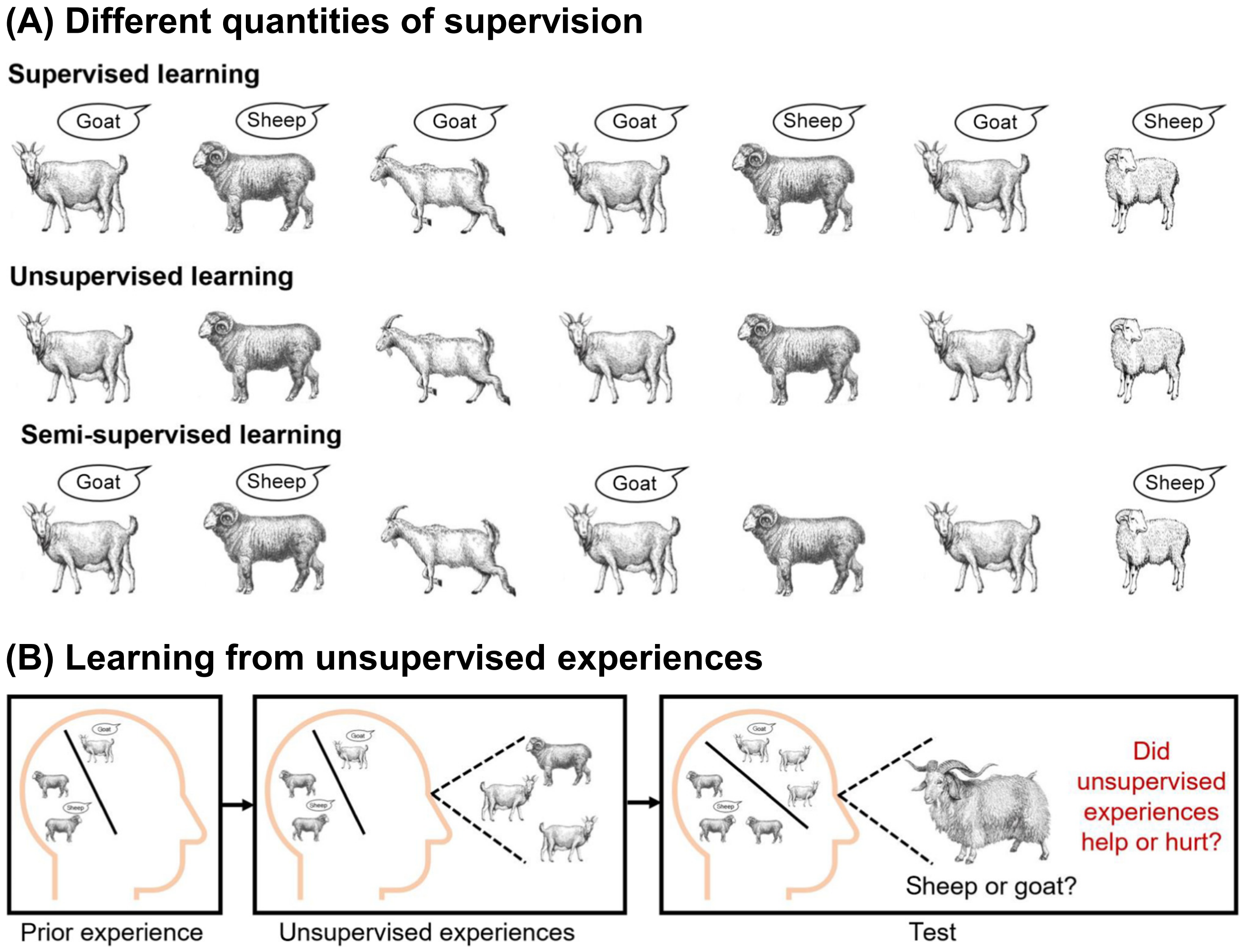
[1]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Gatsby Computational Neuroscience Unit, University College London, London, UK
[3]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
[4]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[5]Department of Psychology, University of Texas at Austin, Austin, TX, US
[6]Department of Experimental Psychology, University College London, London, UK

proposing that humans perform **unsupervised learning** by applying information-processing capabilities that they share with machine learning algorithms [4–6]. A simple and intuitive prediction results from this: if humans share unsupervised information-processing capabilities with machines, and machines show benefits leveraging unsupervised data, then humans should benefit from their unsupervised experience in the same way. That is, humans should be able to recover statistical information from their unsupervised experiences and they should be able to combine it with their rare, supervised experiences.

Paradoxically, this is not supported by the scientific literature. In the most basic learning experiments, humans are not guaranteed to extract statistical information from their unsupervised experiences [7–10] or to boost their supervised learning [11–13]. In fact, unsupervised experiences can reduce performance in category learning [14], language learning [15,16], motor learning [17], and

[7]University of Tübingen, Tübingen, Germany
[8]These authors contributed equally.

*Correspondence:
franziska.broeker.15@ucl.ac.uk
(F. Bröker).



Trends in Cognitive Sciences

Figure 1. Learning with and without supervision. (A) Illustration of supervised, unsupervised, and semi-supervised learning problems. (B) Empirical results conflict as to whether unsupervised experiences improve human performance in unsupervised and semi-supervised learning tasks. We refer to the momentary learning from unsupervised experiences as simply 'unsupervised learning' throughout the text. The reader is encouraged to guess whether the test animal is a sheep or a goat. (While most nonexperts make their sheep/goat predictions based on unreliable features, such as woolliness, the easiest way to tell them apart is by their tails: goats point their tails upward while sheep cannot lift their tails. Thus, the test animal in B is a goat.)

stereotyping [18,19]. Thus, instead of supporting the view that unsupervised experiences help humans in their learning, the literature on laboratory studies is riddled with equivocal results about their benefits. In one experiment, people may need feedback to learn how to distinguish between different visual inputs; in another, they do not [7,20].

These results stem from highly influential experimental designs that have shaped our understanding of how humans extract statistical information. Unsupervised studies often use a simple stimulus–response or passive exposure paradigm. These well-controlled designs are popular because they parallel supervised designs, allowing comparisons. In unsupervised studies, learners predict task-appropriate responses from stimuli without feedback. The statistics in the stimuli are the only information available for learning. Supervised studies are close analogs that provide additional corrective feedback or correct labels, giving learners more information.

Outside the laboratory, human learning operates on a larger scale in terms of data and time. For example, an abundance of additional information can inform learning about sheep and goats, such as separate housing. Learning also serves long-term performance in the world rather than on one specific task. Similarly, modern machine learning solves increasingly large-scale **learning problems**. Given that machine learning algorithms can be flexibly chosen for specific problems, supervised algorithms now solve unsupervised problems by adapting the objective of the **learning task**, as in **self-supervised learning**. Another example is large language models, which learn not by receiving feedback on text they generate, but from predicting words in a sequence. This then serves as a foundation model for further supervised fine-tuning on how to engage in friendly chat with users. These developments increase the complexity in technical approaches and terminology that has yet to be reconciled with human learning inside and outside the laboratory. While an analogy between human and machine unsupervised learning is compelling and often assumed, the devil appears to be in the details.

Here, we mainly focus on laboratory studies that test unsupervised or **semi-supervised learning** using well-controlled, influential designs. Other unsupervised paradigms exist, but are rarer [21]. Our narrower focus ensures that results across various learning contexts are informative about the same learning principles. We refer to momentary learning from unsupervised experiences in experimental tasks as simply ‘unsupervised learning’ to differentiate it from momentary learning with supervisory signals. While focusing on laboratory studies, we also present evidence suggesting unsupervised learning to be limited more generally, because it can worsen performance in machines [3] and human learning outside the laboratory [22]. In fact, telling sheep apart from goats is a task on which many people fail despite recurring exposure (Figure 1B).

The unsupervised snowball effect

How can we explain the mysterious results? When does unsupervised learning help and when does it not? We think that the answer lies in the way in which unsupervised learning is affected by the relationship between the experimenter-defined task and the representations that subjects have acquired from prior experience (**representation-to-task alignment** [14]). Concretely, we propose the unsupervised learning mechanism to be **self-reinforcement**, by which humans learn from their own predictions, such that pre-existing associations between experiences and appropriate responses are strengthened (Figure 2B, Key figure) and decision confidence increases. For example, when seeing the woolly goat in Figure 1B, readers who categorize by woolliness would incorrectly self-reinforce their predictions that it is a sheep, whereas readers who know to attend to the tail would correctly self-reinforce their prediction that it is a goat.

Glossary

Learning algorithm: specific algorithm used to maximize a task objective, which can be supervised or unsupervised. A supervised algorithm (e.g., a standard neural network) learns an input/stimulus to output/response mapping and uses supervision to improve its predictions. Apart from solving supervised learning problems, supervised algorithms can also be used to tackle unsupervised learning problems by adapting the task objective (e.g., self-supervised learning). An unsupervised algorithm (e.g., a standard Bayesian Graphical Model) extracts information from the inputs/stimuli without accessing ground-truth supervision. Unsupervised algorithms are designed to solve unsupervised problems, but can also be adapted to tackle supervised learning problems.

Learning problem: type of learning problem that is partially defined by the data, especially whether supervision is available (i.e., supervised or unsupervised learning problem).

Learning task: specific task (or task objective) that is defined within a learning problem and which can be supervised or unsupervised. In experimental studies, the task objective derives from the experimenter-defined stimulus–response mapping.

Representation-to-task alignment: degree to which the internal representations of a learning system create a similarity space that suggests an input/stimulus to output/response mapping that is in agreement with the objective mapping defined by the task.

Self-reinforcement: mechanism by which a system learns from its own predictions in lieu of ground-truth supervision. This has the effect that existing predictions from inputs/stimuli to outputs/responses are strengthened. In principle, this mechanism can be implemented by both supervised and unsupervised algorithms. This mechanism is popular in semi-supervised machine learning and is called ‘self-training’ or ‘pseudo-labeling’.

Self-supervised learning: machine learning approach that solves an unsupervised learning problem by turning it into a supervised task so that a supervised algorithm can be applied. Since no external supervision is available, supervision is created directly from the unsupervised data.

Semi-supervised learning: learning in a problem/task that offers a mixture of supervised and unsupervised inputs/stimuli.

Supervised learning: learning in a problem/task that requires the learning of an input/stimulus to output/response mapping and in which ground-truth supervision is available.

Supervision/feedback: in machine learning, supervision is defined as the delivery of ground-truth outputs (e.g., labels) following some inputs (e.g., images). In human learning studies, supervision more often refers to the delivery of corrective feedback (e.g., correct/incorrect response) on their response to some preceding stimulus.

Unsupervised learning: learning in a problem/task without supervision, simply through extraction of information from the observation of inputs/stimuli.

Key figure

The unsupervised snowball effect

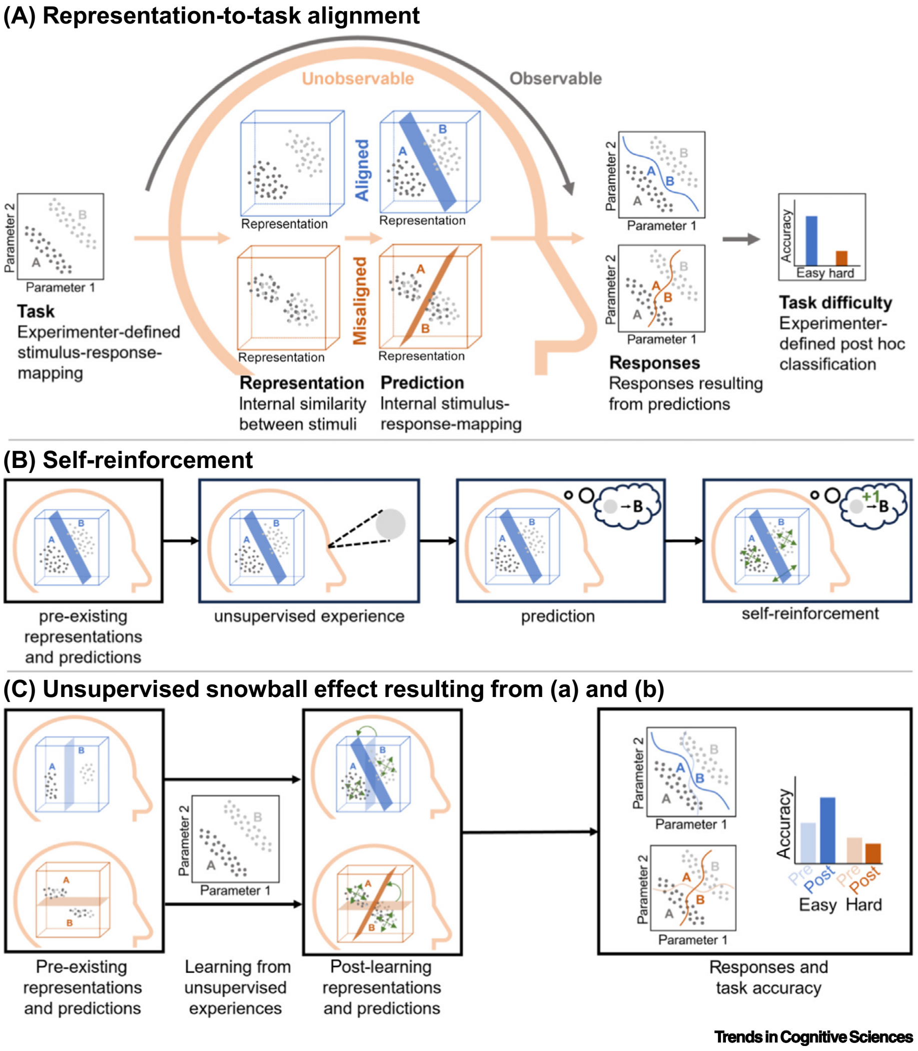

(A) Representation-to-task alignment
Unobservable
Observable
Parameter 2
Parameter 1
A
B
Task
Experimenter-defined stimulus-response-mapping
Aligned
Misaligned
Representation
Representation
Representation
Representation
Representation
Internal similarity between stimuli
Prediction
Internal stimulus-response-mapping
Responses
Responses resulting from predictions
Accuracy
Easy hard
Task difficulty
Experimenter-defined post hoc classification
(B) Self-reinforcement
→B
+1
pre-existing representations and predictions
unsupervised experience
prediction
self-reinforcement
(C) Unsupervised snowball effect resulting from (a) and (b)
Pre-existing representations and predictions
Learning from unsupervised experiences
Post-learning representations and predictions
Accuracy
Pre
Post
Pre
Post
Easy
Hard
Responses and task accuracy
Trends in Cognitive Sciences

Figure 2. Two key factors affect unsupervised learning: representation-to-task alignment and self-reinforcement, resulting in the unsupervised snowball effect, as illustrated in the example of a category learning task. (A) Relationship between experimenter-defined task, its internal representation, and the resulting predictions, responses, and accuracy. Factors including prior experience, context, or attention transform observed stimuli and warp their similarities into an internal representational space that might or might not recover experimenter-defined task statistics. If learners have a task-aligned representation, stimuli from different categories are sufficiently separated in the learner's representational space such that it supports accurate predictions. The task will appear easy, and performance will be high. If learners have a task-misaligned representation, items from different categories are not well separated in the learner's representational space, such that they make incorrect predictions based on whichever task-irrelevant statistics their representations reflect. The task will appear hard, and performance will be low. Thus, we can assume an equivalence between alignment in representations, accuracy of predictions, and task difficulty. (B) Self-reinforcement of predictions. When a stimulus is observed without supervision, an appropriate response is predicted and subsequently self-reinforced. This results in changes in the representations and predictions. (C) If prior representations and predictions are sufficiently aligned with the task, self-reinforcement leads to performance improvement. In the case of misalignment, self-reinforcement has detrimental or no effect on performance. This results in a snowball effect, the course of which can only be changed if supervision is provided to correct mistakes and align representations with the task. Adapted from [19] (B).

Since strengthening predictions snowballs existing learning without changing its course, self-reinforcement can help or hurt depending on how accurate the predictions are for the task at hand (Figure 2A). Self-reinforcing predictions that are largely correct will improve performance in the task. However, predictions will only be largely correct if prior experiences shaped the learner's representations in a way that new experiences elicit appropriate predictions. If this is the case, representations and task are aligned, the task feels 'easy', and supervision is superfluous. By contrast, self-reinforcing predictions that are largely incorrect will have a detrimental, or at best no, effect on performance. Predictions will be largely incorrect if prior experiences have shaped the learner's representations to be misaligned with the task. In this case, the task feels 'hard', and supervision is necessary to adjust the unhelpful representations and predictions of the learner. That self-reinforcing existing representations results in these types of learning dynamics has previously been described in the specific context of unsupervised Hebbian (correlational) learning [23,24]. Our framing of unsupervised learning in terms of representation-to-task alignment and self-reinforcement is more general in that it does not assume specific representations or a specific computational model of learning.

This type of self-reinforcing snowball effect can also be seen when trying to master a new skill, such as playing the violin. This requires practicing with the correct technique because a faulty technique engrains mistakes if left uncorrected. Thus, from our perspective, the equivocal results in the literature about the benefit of unsupervised experiences do not reflect a conflict, but are in fact expected from representation-to-task alignment and its interaction with unsupervised self-reinforcement. Our argument not only follows an intuitive logic, but is also supported by the theoretical principles that allow machine learning algorithms to leverage unsupervised data on many, but not all, occasions (Box 1).

Here, we provide support for this perspective by synthesizing various cognitive science literatures that have long investigated the questions about how feedback influences human learning. The

Box 1. Theoretical principles predict unsupervised snowball effect

We propose that human predictions self-reinforce in the absence of supervision. Since self-reinforcement simply snowballs prior learning, it can help or hurt performance depending on whether predictions and their underlying representations align with the task. Unsupervised learning only succeeds in tasks aligned with the learner's representations.

This intuitive reasoning is supported by the theoretical and computational principles that allow unsupervised and semi-supervised machine learning algorithms to be successful. Inevitably, unsupervised learning can only recover ground-truth structure in the data if this structure is reflected in salient data statistics. For example, for clustering to work, similar points must belong to the same cluster and dissimilar points must belong to different clusters (this is known as the cluster assumption [95]). In other words, clusters need to be sufficiently easy to tell apart to be accurately recovered. In the same way, successful semi-supervised learning requires the to-be-learned input classes to be sufficiently distinctive to work effectively [3,96–98]. Given that this is not always guaranteed and, in practice, is often difficult to validate, unsupervised data are not guaranteed to boost the supervised performance of an algorithm. In fact, much of the success of semi-supervised machine learning could be due to standard-practice data curation, which removes difficult data points from unsupervised training with the effect that input classes become more distinct [99]. Thus, while learning from unlabeled data has led to the much-reported performance boosts in machine learning, it can also lead to degradation. In fact, reports of performance degradation following the addition of unsupervised data exist and are likely under-reported [3].

Returning to empirical studies, sufficient cluster 'distinctiveness' may appear to be a theoretical prerequisite that is easy enough to control experimentally to assess successful, rather than detrimental, unsupervised learning. However, there is a subtle, yet crucial, twist: while experimental tasks may appear to comply with the prerequisite in the experimenter-defined input space, they can simultaneously violate it in the space relevant for learning, which is not routinely assessed: the learner's internal representations of the input space (see Figure 2A in the main text). When overlooked, equivocal results about the benefit of unsupervised experiences can appear conflicting when, in fact, they are predictable. To understand whether results conflict or are simply evidence for the varied directions unsupervised self-reinforcement can take, the alignment between internal representations and experimenter-defined task needs to be considered.

fact that related research fields have largely developed in isolation allows us to test our predictions against their extensive independent evidence. First, we show that representation-to-task alignment correlates with the efficacy of unsupervised learning, as predicted by our hypothesized unsupervised snowball effect. The evidence we consider for the effect of alignment is often somewhat indirect because learners' representations, let alone their alignment with the task, are not typically assessed. Thus, we leverage the equivalences between representation-to-task alignment, predictions, and task difficulty as described in Figure 2A to contextualize the results. Second, we show that unsupervised self-reinforcement has been reported repeatedly across diverse learning settings. We conclude by discussing the implications of our analysis and promising future avenues.

Representation-to-task alignment determines efficacy of unsupervised learning

Representation-to-task alignment is a theoretical concept capturing how well a learner's representations set them up for learning in a new task. Alignment is sufficient when task-relevant statistics are prominent in the representations (e.g., well-separated clusters), when only adaptation of existing representations is needed (e.g., repositioning cluster centers), or both when a beneficial learning sequence builds on prominent representations and subsequently adapts them (e.g., an easy-to-hard curriculum). In these cases, performance is high and tasks are easy (Figure 2A). Given that representation-to-task alignment is independent of any specific type of representation or task, we can expect to observe its effects on the efficacy of unsupervised learning across all types of learning. Here, we test this prediction against the evidence from different, independent literatures.

Perceptual and category learning

Perceptual and category learning experiments share many methodological commonalities. Perceptual learning investigates how perception is changed because of experience with sensory inputs, such as the ability to distinguish different line lengths. This fundamental form of learning is often studied by manipulating simple, physical stimulus dimensions (e.g., line length). Category learning investigates the process of assigning labels (or other distinct responses) to groups of inputs, such as assigning either 'sheep' or 'goat' to each input. This is often studied by manipulating stimulus distributions and boundaries defining categories within them. Stimuli can range from simple shapes or sounds, akin to those used in perceptual learning, to complex, high-dimensional artificial objects. In both paradigms, learners are usually presented with stimuli on a trial-by-trial basis and respond by guessing category membership or, in the case of perceptual learning, by making a same-different judgment between two stimuli.

The perceptual learning literature has extensively studied the effect of different forms of supervision [25,26] and, thus, serves as a superb source of evidence on the effectiveness of unsupervised learning. Results can be summarized simply: unsupervised perceptual training can help in some, but not all, tasks. It does this in a way that correlates with task difficulty, as predicted by our representation-to-task alignment view, which requires sufficient class separation or convenient presentation order. Concretely, unsupervised learning helps if the task is easy and training accuracy is high, as predicted for aligned tasks, [27] or if high-accuracy, easy trials precede or are interleaved with low-accuracy, difficult trials [28–30]. By contrast, feedback appears necessary for learning when task difficulty is high and initial performance is low, as predicted for misaligned tasks [27,31].

Unsupervised and semi-supervised categorization studies in adults echo results from perceptual learning: unsupervised experiences facilitate learning in easier tasks, but not in more difficult ones [9,10]. Learning to separate low-variability categories is easy (aligned task) and equally effective with or without feedback, whereas learning to separate high-variability categories is hard (misaligned task) and requires feedback [32]. Extending this finding, category learning is influenced by the degree of within-category variance [8], with unsupervised learning being most effective and robust when categories are statistically dense and category separation is large [33–35]. This further indicates that sufficient class separation is necessary for successful unsupervised learning (Box 1).

Moreover, unsupervised experiences can have both beneficial and detrimental effects in the exact same task, depending on the alignment of a learner's representations [14]. This pattern is also reflected across tasks. In simple category structures, where stimuli vary along a single dimension, learners can recover categories [20] or shift previously supervised category boundaries without feedback [36–40]. By contrast, in 2D tasks, subjects appear unable to recover categories without feedback [7] and the addition of unsupervised experiences does not boost supervised performance [11,12], except under limiting conditions [41–43]. While experimenter-defined task dimensionality does not imply task difficulty per se, in these experiments, representations required to succeed in the 2D tasks were unmistakably less obvious compared with those required for the 1D tasks. In line with these results, prior knowledge relevant to the task can enhance unsupervised learning [44].

This pattern of results is echoed in language acquisition. When learning non-native phonetic contrasts, unsupervised exposure has been shown to be unsuccessful unless it is complemented by sufficient supervised learning [45] or only involves shifting boundaries of existing phonetic contrasts [46], or if phonetic contrasts are made distinctive [47,48]. We can rephrase these results within our perspective: learning new phonetic contrasts is challenging due to their misalignment with the native speech sound space. To make unsupervised exposure succeed, the task needs to be simplified either by providing feedback that fosters the formation of more aligned representations, by changing the task to only involve modulation of existing, sufficiently aligned representations, or by amplifying the to-be-learned contrast as a form of class separation. Similarly, unsupervised exposure to an artificial language leads to simple word learning, whereas learning its complex syntactic regularities requires feedback [49]. Furthermore, research on infants' capacity to integrate labeled and unlabeled exposure to new categories indicates that learning is successful only when labels are introduced initially, but not when they are presented at the end or omitted entirely [50,51]. This lends credence to our prediction that supervision is required to transition from a misaligned to an aligned representational space before unsupervised experiences can improve performance. A study investigating children's acquisition of linguistic category labels revealed that unsupervised exposure to structured, straightforward labels (regular plural nouns) impaired performance on unstructured, difficult labels (irregular plural nouns) among younger, error-prone children who had not yet mastered the regularities and irregularities. Conversely, it boosted performance among older, more proficient children capable of making adequate predictions [15,16]. This underscores that the outcomes of unsupervised training can vary within the same task, contingent on the learners' representations.

Pre-exposure studies assess the impact of initial unsupervised exposure on later supervised learning and have received independent attention. The effects of pre-exposure vary with category structures [13], with improvements seen for statistically dense categories [52] and exposure to easy stimuli [53,54]. This is in line with our perspective: unsupervised pre-exposure helps in easy tasks but does not affect, or even hinders, difficult ones. Interestingly, rat studies show the opposite (Box 2). This discrepancy is likely due to humans' ability to reason about tasks [55].

Box 2. Results requiring further attention

Pre-exposure in rodents

Interestingly, the effects of unsupervised pre-exposure in rodents are the opposite of those observed in humans. Rodent studies showed that unsupervised learning benefits are greater when stimuli are perceptually similar and, thus, hard to discriminate [100]. Conversely, rodent learning can be hindered when the stimuli are perceptually distinct and, thus, easier to discriminate [101,102]. This effect is attributed to a combination of two learning principles: unsupervised differentiation, which refines representations over time, and latent inhibition, which reduces the associability between inputs and a response [102]. In this context, latent inhibition could explain the slower learning seen after exposure to stimuli that are easily distinguishable.

The opposing effects observed in animals and humans could be due to humans' awareness of their participation in an experiment, leading to heightened attention to stimuli and potential weakening of latent inhibition [55,103]. This is supported by the reversal of pre-exposure effects in rats when using hedonic stimuli, which are believed to stimulate attention [104]. Moreover, interleaving unsupervised and supervised trials in mice appear more effective compared with unsupervised pre-exposure [105], potentially also modulated by attentional factors.

Blocked testing effects

Although understanding learning is important, it is also important to examine how learning could be helped. Across domains, research on optimal training schedules shows that interleaving supervised training with blocks of unsupervised testing consistently improves human learning compared with no testing or restudying of materials. It helps learning and retention of materials preceding or following testing [106,107] and even replacing interim active testing with passive exposure improves performance [108,109]. While individual studies highlight the benefits of supervised testing, particularly its ability to correct inaccuracies and confirm low-confidence predictions [110], a meta-analysis revealed that unsupervised testing benefits are comparable [111]. Taken together, these results appear to suggest that unsupervised testing is exclusively beneficial, a finding that would contradict our unsupervised snowballing theory. However, occasional evidence of performance interactions with learner proficiency and confidence suggest that representation-to-task alignment effects are at play and could simply have gone under-reported.

Selective feedback

Real-world feedback is selective and action dependent, which can lead to learning traps due to unchallenged false predictions [56]. For example, a negative first impression may deter future interactions, preventing the revision of potentially false initial impressions [57]. Similarly, stereotyping can be perpetuated by initial negative experiences with a group, leading to future avoidance. This selective information sampling prevents updating of false predictions about group members, and the likelihood of future avoidance increases when predictions are made without feedback [18]. Consequently, stereotyping intensifies over time, with untested predictions often misremembered as validated [19]. In this way, the selective-feedback literature highlights the detrimental effects of unsupervised learning when predictions are misaligned with reality, as seen in stereotyping.

Expertise

So far, we have seen that unsupervised learning effects vary in controlled laboratory studies. To gauge whether this generalizes to real-world learning, we can assess uncontrolled, long-term learning. Expertise is the product of extensive learning from a varying quantity and quality of supervisory signals outside the laboratory. For instance, radiologists initially receive supervised training but later get less feedback, often not knowing if their diagnoses were correct. If unsupervised experiences had only beneficial effects, we would expect performance to improve over time, leading to expertise even without supervision. However, this prediction has received substantial opposition [58–62] and has even led critics to claim that 'At best, experience is an uncertain predictor of degree of expertise. At worst, experience reflects seniority – and little more.' [60].

Biases, a form of prior expectations, can distort learning and hinder steady improvement through experience. For instance, confirmation bias gives more weight to information that aligns with learners' expectations, skewing learning away from actual evidence [63,64]. In other circumstances, learners may attribute their failure to external factors instead of modifying their erroneous behavior so that performance deteriorates [22].

Irrespective of how expert performance is reached, the expertise literature supports, on a more general level, the claim that unsupervised experiences alone do not guarantee improvement. Instead, reliable improvement appears to require rapid and regular feedback on decisions [62]. Given that acquiring expertise is not easy, but involves learning new skills beyond prior knowledge, these results fit well with our representation-to-task alignment perspective. This is further supported by work showing that initial feedback and guidance are crucial for skill learning [65]. For instance, a laboratory study shows that withdrawing feedback early in motor skill learning, when errors are high (inaccurate predictions), causes performance to deteriorate, whereas doing so later, when errors are low (accurate predictions), enables the skill to be maintained or improved [17].

Self-reinforcement underlies unsupervised learning

While representation-to-task alignment can predict the effectiveness of unsupervised learning, it does not provide a mechanism. Several specific learning procedures have been explored in this context, all of which have self-reinforcement at their core, where learning uses the predictions of the system in lieu of ground-truth supervision, snowballing existing learning without altering its direction.

Perceptual learning, category learning, and expertise

The perceptual learning literature not only supports representation-to-task alignment, but also offers strong evidence for unsupervised self-reinforcement, formalized by Hebbian learning models. Unsupervised Hebbian learning can improve or degrade performance depending on how well representations serve learning a task [23,24]. A Hebbian model that learns from both unsupervised and supervised experiences by adapting representations and their associations with responses [66,67] is successful in accounting for a broad range of results [27]. While trial-by-trial category learning is only rarely modeled, self-reinforcement models have demonstrated their ability to account for semi-supervised categorization [14,68] and can also predict unsupervised learning trajectories in children acquiring linguistic labels [16]. In expertise studies, computational work is limited. However, theories of closed-loop motor skill learning suggest internal estimates guide learning in the absence of feedback leading to either performance gains or decrements [69].

Selective feedback

As described earlier, false predictions that remain unchallenged can, for example, lead to the perpetuation of stereotypes. This can be accounted for by models using unsupervised self-reinforcement [18,19]. Predictions also remain unchallenged when some actions are never followed by feedback (i.e., unsupervised actions). Here, the same self-reinforcement can be observed: humans learn from their own predictions as if they received validation for them (constructivist coding hypothesis [70–72]), which can be modeled by a self-reinforcement mechanism [71].

Internal feedback signals

Self-reinforcement requires internal learning signals independent of external supervision. While the neural mechanisms involved in external supervision (or at least rewards and punishments) are fairly well understood [73], knowledge of the self-generated feedback signals of the brain is limited. Recent studies indicate that brain areas active during external feedback processing are also active when feedback is inferred [74–76]. Moreover, choice consistency and subjective confidence increase in the absence of feedback reflecting self-reinforcement [77], which is in line with evidence that chosen actions carry more internal weight compared with unchosen ones [78]. Subjective rewards can also self-reinforce choices [79]. Large-scale, real-world studies indicate that this can cause people to fall into a learning trap, ceasing exploration and exploiting even when better options exist [80], which an error-driven learning model can account for by aligning subjective preferences with past choices [81]. Neuroimaging also shows that preferences are updated online and only for remembered choices [82]. Moreover, replay, another active research area, involves a form of self-reinforcement in which the brain rehearses past experiences through offline neural reactivation [83,84]. Overall, research supports the use of unsupervised self-reinforcement mechanisms by the brain, with internal signals, such as confidence, having a key role when feedback is absent.

Concluding remarks

In summary, studies across different literatures and learning domains support our perspective: humans self-reinforce their predictions in the absence of supervision, which can either help or hurt performance depending on the alignment between the learner's representations and the task. While we focused on studies testing unsupervised learning under controlled conditions, the expertise literature suggests that these considerations are also relevant to naturalistic settings. This shift in perspective resolves the paradox of predicting learning successes and failures in the laboratory, and fundamentally alters what we expect from unsupervised learning. Unsupervised learning may not be the knight that battles to save us when we lack supervision; instead, it appears to wield a double-edged sword. This raises new questions and lays the foundation for future research on the role of supervision in learning that will have implications for the design of instruction and learning over the lifespan (see Outstanding questions).

A key implication of this perspective is that a deeper understanding of unsupervised learning requires consideration of the alignment between mental representation and task. This is challenging because alignment depends on specific stimuli, task structures, and learners' representations. Efficiently assessing and modeling alignment to account for individual tasks and learners is an important future direction that can build on recent advances [85–88]. In fact, assessing alignment is also important for predicting supervised learning [89,90], memory [91], and perception [92], which suggests that it also applies to naturalistic, large-scale unsupervised learning. Future models need to make explicit the concrete relationship between alignment and learning and be constrained by neural evidence on biologically supported mechanisms [93].

Our efforts to understand when unsupervised learning succeeds and fails have illuminated the rich interconnections between historically separate research areas that can be leveraged in future studies. Beyond the topics discussed here, relevant research also encompasses areas such as attention [94] and training schedules (Box 2). Linking results across these domains promotes a more rigorous examination of learning principles.

Future research should also go beyond the traditional approach of studying unsupervised learning in isolation. To understand why humans manage to learn despite all difficulties, we need to explore how supervised and unsupervised learning mechanisms interact and relate to feedback sources more akin to reinforcement, self-supervised, or sequential learning that are blended in modern machine learning systems. Crucially, future work should explore how unsupervised self-reinforcement and learning from (self-)supervisory signals coexist in humans, who may use one general-purpose mechanism instead of different special-purpose algorithms as machines do. This crosstalk could lead to a more holistic theory of human learning, which is important for understanding real-world learning, such as the acquisition of expertise.

In conclusion, we advocate for an interdisciplinary approach to studying the mechanisms of unsupervised learning and the broader role of supervision, which should integrate representational and neural constraints. This new direction contributes to our understanding of learning

Outstanding questions

What exactly is the quantitative relationship between representation-to-task alignment and learning?How does this relate to different sources of the problem (e.g., poor extraction of relevant features versus good feature extraction, but poor cluster separation)? How does this relate to different timescales (e.g., short-term learning to direct attention versus long-term representational change)?

How much representation-to-task alignment is needed for unsupervised learning to help?

How can we measure representation-to-task alignment? How can we incorporate representation-to-task alignment into computational models of learning?

Does representation-to-task alignment affect supervised and unsupervised learning differently?

How is self-reinforcement implemented by the brain? Which role does meta-cognition have in this? Is it affected by brain development?

Does self-reinforcement also affect supervised learning?

How do supervised and unsupervised learning interact? Are they fundamentally different or can they be unified?

How does learning from other feedback signals, such as reward, compare with supervised and unsupervised learning?

How does unsupervised learning compare in humans and animals? Are there differences between implicit/subconscious and deliberate/conscious unsupervised learning?

Which other factors related to the presence and absence of supervision, such as motivation, affect learning?

How does the sequential order (e.g., blocked supervised and unsupervised exposure) affect unsupervised learning?

fundamentals and can improve the design of instructional systems that better support learning across the lifespan to prevent us from mistaking goats for sheep with ever greater confidence.

Acknowledgments

This work was supported by the Gatsby Charitable Foundation (F.B.); Neuroscience Institute of Carnegie Mellon University (F.B.); the Max Planck Society (F.B., P.D.); the Alexander von Humboldt Foundation (P.D.); ESRC (ES/W007347/1) and a Royal Society Wolfson Fellowship (18302) to B.C.L.; and NSF BCS2420979 and NSF BCS2346989 to L.L.H. Some of this work has also been described in the PhD thesis of the first author (F. Bröker, PhD thesis, University College London, 2022).

Declaration of interests

None declared by authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Microsoft Copilot in order to shorten the text and improve the writing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Tian, K. *et al.* (2017) Deepcluster: a general clustering framework based on deep learning. *Lect. Notes Comput. Sci.* 10535, 809–825
2. Devlin, J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Burstein, J. *et al.*, eds), pp. 4171–4186, Association for Computational Linguistics
3. Van Engelen, J.E. and Hoos, H.H. (2020) A survey on semi-supervised learning. *Mach. Learn.* 109, 373–440
4. Lee, T.S. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *JOSA A* 20, 1434–1448
5. Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719
6. Tenenbaum, J.B. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285
7. Ashby, F.G. *et al.* (1999) On the dominance of unidimensional rules in unsupervised categorization. *Percept. Psychophys.* 61, 1178–1199
8. Ell, S.W. *et al.* (2012) Unsupervised category learning with integral-dimension stimuli. *Q. J. Exp. Psychol.* 65, 1537–1562
9. Wade, T. and Holt, L.L. (2005) Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* 118, 2618–2633
10. Emberson, L.L. *et al.* (2013) Is statistical learning constrained by lower level perceptual organization? *Cognition* 128, 82–102
11. Vandist, K. *et al.* (2009) Semisupervised category learning: the impact of feedback in learning the information-integration task. *Atten. Percept. Psychophysiol.* 71, 328–341
12. McDonnell, J.V. *et al.* (2012) Sparse category labels obstruct generalization of category membership. in *Proc. Annu. Meet. Cogn. Sci. Soc.* 34, 749–754
13. Wills, A.J. *et al.* (2004) The role of category structure in determining the effects of stimulus preexposure on categorization accuracy. *Q. J. Exp. Psychol. Sect. B* 57, 79–88
14. Bröker, F. *et al.* (2022) When unsupervised training benefits category learning. *Cognition* 221, 104984
15. Ramscar, M. and Yarlett, D. (2007) Linguistic self-correction in the absence of feedback: a new approach to the logical problem of language acquisition. *Cogn. Sci.* 31, 927–960
16. Ramscar, M. *et al.* (2013) Error and expectation in language learning: The curious absence of mouses in adult speech. *Language* 89, 760–793
17. Newell, K.M. (1974) Knowledge of results and motor learning. *J. Mot. Behav.* 6, 235–244
18. Allidina, S. and Cunningham, W.A. (2021) Avoidance begets avoidance: a computational account of negative stereotype persistence. *J. Exp. Psychol. Gen.* 150, 2078–2099
19. Cox, W.T.L. *et al.* (2022) Untested assumptions perpetuate stereotyping: learning in the absence of evidence. *J. Exp. Soc. Psychol.* 102, 104380
20. Rosenthal, O. *et al.* (2001) Forming classes by stimulus frequency: behavior and theory. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4265–4270
21. Love, B.C. (2002) Comparing supervised and unsupervised category learning. *Psychon. Bull. Rev.* 9, 829–835
22. Kc, D. *et al.* (2013) Learning from my success and from others' failure: evidence from minimally invasive cardiac surgery. *Manag. Sci.* 59, 2435–2449
23. McClelland, J.L. (2001) Failures to learn and their remediation: a Hebbian account. In *Mechanisms of Cognitive Development* (McClelland, J.L. and Steigler, R., eds), pp. 109–134, Psychology Press
24. McClelland, J.L. (2006) How far can you go with Hebbian learning, and when does it lead you astray. *Process. Change Brain Cogn. Dev. Atten. Perform. XXI* 21, 33–69
25. Dosher, B.A. and Lu, Z.-L. (2009) Hebbian reweighting on stable representations in perceptual learning. *Learn. Percept.* 1, 37–58
26. Dosher, B.A. and Lu, Z.-L. (2017) Visual perceptual learning and models. *Annu. Rev. Vis. Sci.* 3, 343–363
27. Liu, J. *et al.* (2010) Augmented Hebbian reweighting: interactions between feedback and training accuracy in perceptual learning. *J. Vis.* 10, 29
28. Ahissar, M. and Hochstein, S. (1997) Task difficulty and the specificity of perceptual learning. *Nature* 387, 401–406
29. Liu, J. *et al.* (2012) Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vis. Res.* 61, 15–24
30. Asher, J.M. and Hibbard, P.B. (2020) No effect of feedback, level of processing or stimulus presentation protocol on perceptual learning when easy and difficult trials are interleaved. *Vis. Res.* 176, 100–117
31. Shiu, L.-P. and Pashler, H. (1992) Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Percept. Psychophys.* 52, 582–588
32. Homa, D. and Cultice, J.C. (1984) Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 83
33. Kloos, H. and Sloutsky, V.M. (2008) What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *J. Exp. Psychol. Gen.* 137, 52
34. Pothos, E.M. *et al.* (2011) Measuring category intuitiveness in unconstrained categorization tasks. *Cognition* 121, 83–100

35. Vong, W.K. *et al.* (2016) The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychon. Bull. Rev.* 23, 230–238
36. Zhu, X. *et al.* (2007) Humans perform semi-supervised classification too. *Proc. AAAI Conf. Artif. Intell.* 22, 864–870 AAAI
37. Lake, B. and McClelland, J. (2011) Estimating the strength of unlabeled information during semi-supervised learning. *Proc. Annu. Meet. Cogn. Sci. Soc.* 33, 1400–1405
38. Kalish, C.W. *et al.* (2011) Can semi-supervised learning explain incorrect beliefs about categories? *Cognition* 120, 106–118
39. Kalish, C.W. *et al.* (2015) Drift in children's categories: when experienced distributions conflict with prior learning. *Dev. Sci.* 18, 940–956
40. Gibson, B.R. *et al.* (2015) What causes category-shifting in human semi-supervised learning? *Proc. Annu. Meet. Cogn. Sci. Soc.* 37, 794–799
41. Rogers, T. *et al.* (2010) Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. *Proc. Annu. Meet. Cogn. Sci. Soc.* 32, 2320–2325
42. Rogers, T. *et al.* (2010) Humans learn using manifolds, reluctantly. *Adv. Neural Inf. Proces. Syst.* 23, 1–9
43. Vandist, K. *et al.* (2019) Semisupervised category learning facilitates the development of automaticity. *Atten. Percept. Psychophysiol.* 81, 137–157
44. Clapper, J.P. (2007) Prior knowledge and correlational structure in unsupervised learning. *Can. J. Exp. Psychol. Rev. Can. Psychol. Expérimentale* 61, 109–127
45. Wright, B.A. *et al.* (2019) Semi-supervised learning of a nonnative phonetic contrast: How much feedback is enough? *Atten. Percept. Psychophysiol.* 81, 927–934
46. Chládková, K. *et al.* (2022) Unattended distributional training can shift phoneme boundaries. *Biling. Lang. Cogn.* 25, 827–840
47. 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
48. Escudero, P. *et al.* (2011) Enhanced bimodal distributions facilitate the learning of second language vowels. *J. Acoust. Soc. Am.* 130, EL206–EL212
49. Frinsel, F.F. *et al.* (2024) The role of feedback in the statistical learning of language-like regularities. *Cogn. Sci.* 48, e13419
50. LaTourrette, A. and Waxman, S.R. (2019) A little labeling goes a long way: semi-supervised learning in infancy. *Dev. Sci.* 22, e12736
51. LaTourrette, A. and Waxman, S.R. (2022) Sparse labels, no problems: Infant categorization under challenging conditions. *Child Dev.* 93, 1903–1911
52. Unger, L. and Sloutsky, V.M. (2022) Ready to learn: incidental exposure fosters category learning. *Psychol. Sci.* 33, 999–1019
53. Milton, F. *et al.* (2014) The effect of pre-exposure on family resemblance categorization for stimuli of varying levels of perceptual difficulty. *Proc. Annu. Meet. Cogn. Sci. Soc.* 36, 1018–1023
54. Milton, F. *et al.* (2020) The effect of preexposure on overall similarity categorization. *J. Exp. Psychol. Anim. Learn. Cogn.* 46, 65–82
55. Angulo, R. *et al.* (2019) Stimulus comparison: effects of the preexposure schedule and instructions for perceptual learning and attention. *Learn. Motiv.* 65, 20–32
56. Rich, A.S. and Gureckis, T.M. (2018) The limits of learning: exploration, generalization, and the development of learning traps. *J. Exp. Psychol. Gen.* 147, 1553
57. Denrell, J. (2005) Why most people disapprove of me: experience sampling in impression formation. *Psychol. Rev.* 112, 951
58. Brehmer, B. (1980) In one word: not from experience. *Acta Psychol.* 45, 223–241
59. Garb, H.N. (1989) Clinical judgment, clinical training, and professional experience. *Psychol. Bull.* 105, 387
60. Shanteau, J. *et al.* (2002) Performance-based assessment of expertise: How to decide if someone is an expert or not. *Eur. J. Oper. Res.* 136, 253–263
61. Ericsson, K.A. (2004) Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad. Med.* 79, S70–S81
62. Kahneman, D. and Klein, G. (2009) Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64, 515
63. Wason, P.C. (1960) On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140
64. Nickerson, R.S. (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220
65. Dunphy, B.C. and Williamson, S.L. (2004) In pursuit of expertise. Toward an educational model for expertise development. *Adv. Health Sci. Educ.* 9, 107–127
66. Petrov, A.A. *et al.* (2005) The dynamics of perceptual learning: an incremental reweighting model. *Psychol. Rev.* 112, 715
67. Petrov, A.A. *et al.* (2006) Perceptual learning without feedback in non-stationary contexts: data and model. *Vis. Res.* 46, 3177–3197
68. Gibson, B.R. *et al.* (2013) Human semi-supervised learning. *Top. Cogn. Sci.* 5, 132–172
69. Adams, J.A. (1971) A closed-loop theory of motor learning. *J. Mot. Behav.* 3, 111–150
70. Elwin, E. *et al.* (2007) Constructivist coding: learning from selective feedback. *Psychol. Sci.* 18, 105–110
71. Henriksson, M.P. *et al.* (2010) What is coded into memory in the absence of outcome feedback? *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1–16
72. Elwin, E. (2013) Living and learning: reproducing beliefs in selective experience: living and learning. *J. Behav. Decis. Mak.* 26, 327–337
73. Schultz, W. (2007) Behavioral dopamine signals. *Trends Neurosci.* 30, 203–210
74. Daniel, R. and Pollmann, S. (2012) Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* 59, 3457–3467
75. Guggenmos, M. *et al.* (2016) Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* 5, e13388
76. Rouault, M. *et al.* (2019) Forming global estimates of self-performance from local confidence. *Nat. Commun.* 10, 1141
77. Ptasczynski, L.E. *et al.* (2022) The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLoS Comput. Biol.* 18, e1010580
78. Sakamoto, Y. and Miyoshi, K. (2024) A confidence framing effect: flexible use of evidence in metacognitive monitoring. *Conscious. Cogn.* 118, 103636
79. Vinckier, F. *et al.* (2019) Sour grapes and sweet victories: how actions shape preferences. *PLoS Comput. Biol.* 15, e1006499
80. Riefer, P.S. *et al.* (2017) Coherency-maximizing exploration in the supermarket. *Nat. Hum. Behav.* 1, 0017
81. Hornsby, A.N. and Love, B.C. (2020) How decisions and the desire for coherency shape subjective preferences over time. *Cognition* 200, 104244
82. Voigt, K. *et al.* (2019) Hard decisions shape the neural coding of preferences. *J. Neurosci.* 39, 718–726
83. 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
84. Barry, D.N. and Love, B.C. (2022) A neural network account of memory replay and knowledge consolidation. *Cereb. Cortex* 33, 83–95
85. Houlsby, N.M. *et al.* (2013) Cognitive tomography reveals complex, task-independent mental representations. *Curr. Biol.* 23, 2169–2175
86. Hebart, M.N. *et al.* (2020) Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* 4, 1173–1185
87. Ma, W.J. and Peters, B. (2020) A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv*, Published online May 2, 2020. http://dx.doi.org/10.48550/arXiv.2005.02181
88. Roads, B.D. and Love, B.C. (2021) Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3547–3557, IEEE
89. Aho, K. *et al.* (2022) System alignment supports cross-domain learning and zero-shot generalisation. *Cognition* 227, 105200

90. Roark, C.L. *et al.* (2022) A neural network model of the effect of prior experience with regularities on subsequent category learning. *Cognition* 222, 104997
91. Schurgin, M.W. *et al.* (2020) Psychophysical scaling reveals a unified theory of visual memory strength. *Nat. Hum. Behav.* 4, 1156–1172
92. Zaman, J. *et al.* (2021) Perceptual variability: Implications for learning and generalization. *Psychon. Bull. Rev.* 28, 1–19
93. Golub, M.D. *et al.* (2018) Learning by neural reassociation. *Nat. Neurosci.* 21, 607–616
94. Hammer, R. *et al.* (2015) Feature saliency and feedback information interactively impact visual category learning. *Front. Psychol.* 6, 74
95. Chapelle, O. *et al.* (2006) *Semi-Supervised Learning*, MIT Press
96. Singh, A. *et al.* (2008) Unlabeled data: now it helps, now it doesn't. *Adv. Neural Inf. Proces. Syst.* 21, 1–8
97. Zhu, X. and Goldberg, A.B. (2009) *Introduction to Semi-Supervised Learning*, Springer International Publishing
98. Oymak, S. and Gulcu, T.C. (2020) Statistical and algorithmic insights for semi-supervised learning with self-training. *arXiv*, Published online June 19, 2020. http://dx.doi.org/10.48550/arXiv.2006.11006
99. Ganev, S. and Aitchison, L. (2021) Semi-supervised learning objectives as log-likelihoods in a generative model of data curation. *arXiv*, Published online August 13, 2020. http://dx.doi.org/10.48550/arXiv.2008.05913
100. Oswalt, R.M. (1972) Relationship between level of visual pattern difficulty during rearing and subsequent discrimination in rats. *J. Comp. Physiol. Psychol.* 81, 122
101. Chamizo, V.D. and Mackintosh, N. (1989) Latent learning and latent inhibition in maze discriminations. *Q. J. Exp. Psychol.* 41, 21–31
102. Saksida, L.M. (1999) Effects of similarity and experience on discrimination learning: a nonassociative connectionist model of perceptual learning. *J. Exp. Psychol. Anim. Behav. Process.* 25, 308
103. Graham, S. and McLaren, I. (1998) Retardation in human discrimination learning as a consequence of pre-exposure: latent inhibition or negative priming? *Q. J. Exp. Psychol. Sect. B* 51, 155–172
104. Sanjuán, M. del C. *et al.* (2014) An easy-to-hard effect after nonreinforced preexposure in a sweetness discrimination. *Learn. Behav.* 42, 209–214
105. Schmid, C. *et al.* (2024) Passive exposure to task-relevant stimuli enhances categorization learning. *eLife* 12, RP88406
106. Lee, H.S. and Ahn, D. (2018) Testing prepares students to learn better: the forward effect of testing in category learning. *J. Educ. Psychol.* 110, 203–217
107. Yang, C. and Shanks, D.R. (2018) The forward testing effect: interim testing enhances inductive learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 485–492
108. Wright, B.A. *et al.* (2010) Enhancing perceptual learning by combining practice with periods of additional sensory stimulation. *J. Neurosci.* 30, 12868–12877
109. Wright, B.A. *et al.* (2015) Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. *J. Acoust. Soc. Am.* 138, 928–937
110. Wang, L. and Yang, J. (2021) Effect of feedback type on enhancing subsequent memory: interaction with initial correctness and confidence level. *PsyCh J.* 10, 751–766
111. Adesope, O.O. *et al.* (2017) Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701