Cerebral Cortex, 2022, **00**, 1–13

https://doi.org/10.1093/cercor/bhac054

Original Article

A neural network account of memory replay and knowledge consolidation

Daniel N. Barry, PhD[1,*], Bradley C. Love, PhD[1,2]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H0AP, UK,
[2]The Alan Turing Institute, 96 Euston Road, London NW12DB, UK
*Corresponding author: Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H0AP, UK. Email: daniel.barry@ucl.ac.uk

Abstract

Replay can consolidate memories through offline neural reactivation related to past experiences. Category knowledge is learned across multiple experiences, and its subsequent generalization is promoted by consolidation and replay during rest and sleep. However, aspects of replay are difficult to determine from neuroimaging studies. We provided insights into category knowledge replay by simulating these processes in a neural network which approximated the roles of the human ventral visual stream and hippocampus. Generative replay, akin to imagining new category instances, facilitated generalization to new experiences. Consolidation-related replay may therefore help to prepare us for the future as much as remember the past. Generative replay was more effective in later network layers functionally similar to the lateral occipital cortex than layers corresponding to early visual cortex, drawing a distinction between neural replay and its relevance to consolidation. Category replay was most beneficial for newly acquired knowledge, suggesting replay helps us adapt to changes in our environment. Finally, we present a novel mechanism for the observation that the brain selectively consolidates weaker information, namely a reinforcement learning process in which categories were replayed according to their contribution to network performance. This reinforces the idea of consolidation-related replay as an active rather than passive process.

Key words: consolidation; learning; memory; network; replay.

Introduction

Memory consolidation-related replay

Memory replay refers to the reactivation of experience-dependent neural activity during resting periods. First observed in rodent hippocampal cells during sleep (Wilson and McNaughton 1994), the phenomenon has since been detected in humans during rest (Tambini and Davachi 2013; Hermans et al. 2017; Schapiro et al. 2018; Liu et al. 2019; Wittkuhn and Schuck 2021) and sleep (Schönauer et al. 2017; Zhang et al. 2018). These investigations have revealed replayed experiences are more likely to be subsequently remembered; therefore, replay has been proposed to strengthen the associated neural connections and to protect memories from being forgotten. Replay which supports memory consolidation can be viewed as distinct from task-related replay, the neural reactivation observed during task performance which supports cognitive processes such as memory recall (Jafarpour et al. 2014; Michelmann et al. 2019; Wimmer et al. 2020), visual understanding (Schwartenbeck et al. 2021), decision-making (Liu et al. 2021), planning (Momennejad et al. 2018), and prediction (Ekman et al. 2017). While traditional perspectives view memory consolidation as a gradual process of fixation, whereby memories are stabilized (Squire and Alvarez 1995; McGaugh 2000), in this paper, we advocate the more contemporary view that offline consolidation-related replay is more dynamic in nature (Mattar and Daw 2018). Using a computational approach, we test hypotheses that offline replay may be a creative process to serve future goals, that it matters exactly where in the brain replay occurs, that it helps us at particular stages of learning, and that the brain might actively choose the optimal experiences to replay.

Generative replay of category knowledge

Neural replay which supports memory consolidation during rest and sleep has been traditionally assumed to be veridical such that we commit the events of that day to long-term memory by replaying the episodes as they were originally experienced. However, there are circumstances in which this may be suboptimal or impractical. For example, a desirable outcome of category knowledge consolidation is to generalize to new experiences rather than recognize past instances. This phenomenon has been observed after sleep in infants (Gómez et al. 2006; Friedrich et al. 2015; Horváth et al. 2016) and in adults (Lau et al. 2011). Sleep also

Received: August 23, 2021. **Revised:** January 25, 2022. **Accepted:** January 26, 2022

recovers the generalization of phonological categories (Fenn et al. 2003), preserves generalization performance in perceptual category learning (Graveline and Wamsley 2017), and assists in the abstraction of gist-like prototype representations (Lutz et al. 2017). It is still not understood how the brain consolidates and replays memory in the service of generalization. In addition, although sleep benefits category learning for a limited number of well-controlled experimental stimuli (Schapiro et al. 2017), in the real world, category learning takes place over many thousands of experiences, and storing each individual experience for replay is an impractical proposition. For these reasons, we propose the replay of novel, prototypical category instances would be a more efficient and effective solution. In fact, given the role of the hippocampus in both replay (Zhang et al. 2018) and the generation of prototypical concepts (Hassabis et al. 2007), we consider this as the most likely form of category replay. While evidence for such generative replay of category knowledge has yet to be discovered in the human brain, replay of sequences immediately following task performance in humans has been shown to be flexible in that items can be reordered based on previously learned rules (Liu et al. 2019). This is reminiscent of "preplay" observed during task performance in rodents, where hippocampal "place cells" observed to fire in specific locations reactivate in a different order to represent a route which has not been taken before (Gupta et al. 2010).

Drawing inspiration from these observations, here, we test the idea that replay which facilitates memory consolidation, occurring over extended offline time periods including sleep, might also be generative in nature and that its flexibility may not just apply to the reorganization of learned sequences but to the creation of entirely new instances of a category. While decoding the reordering of stimuli or route knowledge from brain data during replay has been shown to be a tractable approach, detecting entirely new instances of complex categories from the brain represents a significant challenge and has yet to be demonstrated.

One approach to address this question is to simulate these processes in an artificial neural network. Prior research with artificial neural networks has modeled the replay of generated image stimuli (van de Ven et al. 2020). While revealing a promising avenue of investigation, the results of this study cannot be easily extrapolated to the brain or human visual experience. For example, the structure of only 5 convolutional layers in the network employed represents just a fraction of the size of larger models which have been shown to extract visual representations similar in nature to those processed by the brain (Schrimpf et al. 2018), whose complex structure can be compared to the ventral visual stream processing pathway, indicating a possible correspondence in functional architecture (Khaligh-Razavi and Kriegeskorte 2014; Güçlü and van Gerven 2015; Devereux et al. 2018), and whose object recognition performance approaches that of humans (He et al. 2015). Further, the networks employed by van de Ven et al. (2020) had limited visual experience, having been pretrained on just 10 categories of objects. By contrast, an adult human brain will harbor a lifetime of visual knowledge which facilitates the learning of novel concepts. Therefore, to simulate the learning and generative replay of new categories realistically in adults, using an experienced network which contains a preexisting vast body of knowledge about a range of other categories is an essential starting point. Another feature of the aforementioned study which limits the comparison to humans is that the stimuli used were low-resolution photographs measuring 32 × 32 pixels, which do not reflect the complexity of human visual experience. To accurately simulate human learning and replay, much larger, high-resolution images which go some way toward approaching the complexity and richness of everyday human visual experience are required as training stimuli. Finally, prior attempts at replay in neural networks, whether generative (Kemker and Kanan 2017; van de Ven et al. 2020) or veridical (Hayes et al. 2021), have been deployed to address the "catastrophic forgetting" problem—the tendency of artificial networks to forget old categories when new ones are learned (Robins 1995; French 1999). While this has been proposed as a potential mechanism for why biological agents do not suffer from catastrophic forgetting, empirical evidence in support of this hypothesis has not been forthcoming to date. In addition, other solutions have been put forward on how brains and models may avoid catastrophic interference, such as Adaptive Resonance Theory (Grossberg 2013) and elastic weight consolidation (Kirkpatrick et al. 2017).

In this study, we investigated whether offline generative replay of novel concepts facilitated subsequent generalization to new experiences using models which attempt to simulate the human brain and approximate more closely the visual environment in which it learns. To do this, we implemented generative replay in a well-studied deep convolutional neural network (DCNN), which consists of a complex architecture organized into 5 blocks of convolutional layers and boasts of a high "brain-score," indicating the representations it extracts bear a similarity to those extracted by the brain, and it performs favorably to humans in a categorization task (Schrimpf et al. 2018). The network had prior experience of learning 1,000 diverse categories of objects from over a million high-resolution complex naturalistic images, a process which is the network equivalent of a lifetime of visual experience and which yields within the model an optimized, high-functioning visual system. We tasked the model with learning 10 novel categories it had not seen before using similarly high-resolution naturalistic images to those it has seen before, with an average resolution of around 400 × 350 pixels (Deng et al. 2009), representing an approximate 140-fold increase in visual details from stimuli used in prior work. A comparable learning experience in humans would be coming across

10 new categories we had not seen before and using our lifelong experience in processing visual information to extrapolate the relevant identifying features. After learning periods, we then simulated generative replay in the network, which attempted to mimic human consolidation during sleep and monitored the network's performance when it "woke up" the next day to ascertain if we could provide computational support for the theory that such a process underlies the overnight improvements in generalization observed in humans.

Effective neural loci of replay

Another outstanding question regarding replay is, despite being associated with subsequent memory (Zhang et al. 2018), it is not clear where in the brain replay makes a demonstrable contribution toward generalization. Replay has been observed throughout the brain, early in the ventral visual stream (Ji and Wilson 2007; Deuker et al. 2013; Wittkuhn and Schuck 2021), in the ventral temporal cortex (Tambini et al. 2010; de Voogd et al. 2016), the medial temporal lobe (Staresina et al. 2013; Schapiro et al. 2018) the amygdala, (Girardeau et al. 2017; Hermans et al. 2017), motor cortex (Eichenlaub et al. 2020), and prefrontal cortex (Peyrache et al. 2009). It is not known if replay in lower-level brain regions actually contributes to the observed memory improvements or whether the key neural changes are made in more advanced areas, and this question cannot be answered using current neuroimaging approaches. One prior study has implemented replay within an artificial neural network from a single location at the end of the network (van de Ven et al. 2020). However, because the compact architecture of this network did not have a clear functional correspondence with information processing pathways in the brain, and because replay from other locations within the network was not also implemented for comparison, it is difficult to yield predictions from these results regarding effective replay locations in the human brain. In the current study, because we simulated replay in a neural network which processes images in a manner reflective of the human ventral visual stream, we could compare the effectiveness of replay from different layers with a purported representational correspondence to specific regions in the brain. In doing so, we aimed to make predictions about the effective cortical targets of offline memory consolidation in humans.

A time-dependent role for replay

Another open question regarding human replay is the duration of its involvement throughout the learning of novel concepts. It can take humans years to learn and consolidate semantic or conceptual knowledge (Manns et al. 2003), but neuroimaging studies of replay are limited to a time-span of a day or 2, therefore it is still not known how long replay contributes to this process. Humans are thought to "reconsolidate" information every time it is retrieved (Dudai 2012), suggesting replay might play a continual role in the lifespan of memory. However, recordings in rodents have shown that replay diminishes with repeated exposure to an environment over multiple days (Giri et al. 2019), suggesting the brain may only replay recently learned, vulnerable information. Answering this question in humans remains a challenge because of the impracticalities of tracking replay events for extended periods. Simulation in a human-like neural network represents a possible alternative to predict the relative contribution of replay to consolidation over long time periods, an approach which has not been attempted to date. Here, we interleaved daily learning with nights of offline replay in a neural network, which simulates the brain to understand at what stage in learning replay may be most effective in humans.

Replay of weakly learned knowledge

An additional poorly understood principle of replay which we investigated in this study is why consolidation tends to selectively benefit weakly learned over well-learned information (Kuriyama et al. 2004; Drosopoulos et al. 2007; McDevitt et al. 2015; Schapiro et al. 2018). Here, we modeled a candidate mechanism for how this occurs in the brain by adding an auxiliary model (theoretically analogous to the hippocampus) to the neocortical-like model, which could autonomously learn the best consolidation strategy, determining what to replay and when.

Hypotheses

In addressing these outstanding questions regarding replay in the brain, we made a number of predictions. Because earlier brain regions are thought to extract equivalent basic features from all categories, we predicted replay of experience would be more effective in promoting learning at advanced stages of the network. We hypothesized the replay of "imagined" prototypical replay events would be as effective as veridical replay in helping us to generalize to new, unseen experiences, thus supporting our conceptualization of replay as a creative process. We predicted that the benefits of replay may be confined to early in the learning curve when novel category knowledge is being acquired. Finally, we hypothesized that a dynamic interaction between hippocampal- and neocortical-like models would result in the prioritization of weakly learned items, which is in line with behavioral studies of memory consolidation.

Materials and methods

Neural network

To simulate the learning of novel concepts in the brain and to test a number of hypotheses regarding replay, we trained a DCNN on 10 new categories of images. The neural network was VGG-16 (Simonyan and Zisserman 2014). This network is trained on a vast dataset of 1.3 million high-resolution complex naturalistic photographs known as the ImageNet database (Deng et al. 2009), which contains recognizable objects from 1,000 categories in different contexts. The network learns to associate the visual features of an object with its category label until it can recognize examples of that object which it has never seen before, simulating the human ability to generalize prior knowledge to new situations. The network takes a photograph's pixels as input and sequentially transforms this input into more abstract features. It learns to perform these transformations by adjusting 138,357,544 connection weights across many layers. Its convolutional architecture reduces the number of possible training weights by searching for informative features in any area of the photographs.

In these experiments, we task the VGG-16 network with learning 10 new categories of images. To do this, we retained the pretrained "base" of this network, which consisted of 19 layers, organized into 5 convolutional blocks. Within each block, there were convolutional layers and a pooling layer with nonlinear activation functions. To this base, we attached 2 fully connected layers, each followed by a "dropout" layer, which randomly zeroed out 50% of units to prevent overfitting to the training set (Srivastava et al. 2014). At the end of the network, a SoftMax layer was attached, which contained just 10 outputs rather than the original 1,000 and predicted which of 10 classes an image belonged to. To facilitate the learning of 10 new classes, the weights of layers attached to the pretrained base were randomly initialized. All model parameters were free to be trained. In total, 10 new models were trained, each learning 10 new and different classes.

Stimuli

Photographic stimuli for new classes were drawn randomly from the larger ImageNet 2011 fall database (Russakovsky et al. 2015) and were screened manually by the experimenter to exclude classes which bore a close resemblance to classes which VGG-16 was originally trained on. In total, 100 new classes were selected and were randomly assigned to the 10 different models to be trained. Within each class, a set of 1,170 training images, 130 validation images, and 50 test images were selected. The list of the selected classes is available in Supplementary Table 1.

Baseline training

We first trained a model without implementing replay to serve as a baseline measure of network performance and compared with other conditions which implemented replay. Ten models were trained on 10 new and different classes. To further prevent overfitting to the training set, images were augmented before each training epoch. This is similar to a human viewing an object at different locations, or from different angles, and facilitates the extraction of useful features rather than rote memorization of experience. Augmentation could include up to 20° rotation, 20% vertical or horizontal shifting, 20% zoom, and horizontal flipping. Any blank portions of the image following augmentation were filled with a reflection of the existing image. Images were then preprocessed in accordance with Simonyan and Zisserman (2014). Depending on the experiment, the network was trained for 10 or 30 epochs. We used the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.0003. A small learning rate was chosen to reflect the fact that learning new categories in an adult human reflects a "fine-tuning" of an already highly trained visual system. The training batch size was set to 36. The training objective was to minimize the categorical cross-entropy loss over the 10 classes. Training parameters were optimized based on validation set performance. We report the model's performance metrics from the test set only. This is a collection of novel images from each category which the network is tasked with classifying. The network has not previously learned these images nor had its parameters tuned based on them. Therefore performance on these images reflects the model's ability to generalize to new stimuli after training, and is thus termed "generalization performance" in the figures. Training was performed using TensorFlow version 2.2.

Replay

Replay was conducted between training epochs to simulate "days" of learning and "nights" of offline consolidation. We conceptualized replay representations as generative, in other words, they represented a prototype of that category never seen before from a particular point in the network. To generate these representations, the network activations induced by the training images from the preceding epoch were extracted from a particular layer in the network using the Keract toolbox (Remy 2020). For each class separately, a multivariate distribution of activity was created from these activations using the SciPy toolbox (https://scipy.org/). This multivariate normal distribution is an extension of the 1D normal distribution to higher dimensions and is specified by its mean and covariance matrix. This resulted in a single unique distribution for each specific class, which represented the relationship between units of the layer which had been previously observed for that class. We then sampled randomly from this distribution, creating novel activation patterns for that class at that point in the network (Fig. 1A). These novel activation patterns represented a prototype of that category. The end result was a representation that was a rough approximation of the layer's representations of that category if a real image was processed but was novel in nature (see Supplementary Fig. 1). The human brain equivalent would be the approximate pattern of neural activity which is representative of that category at a particular stage in the ventral visual stream. In the brain, these hypothetical prototypical concepts would be likely generated from more high-level regions such as the hippocampus and prefrontal cortex (Hassabis et al. 2007; Bowman et al. 2020). Our model was generative as it could create new samples;

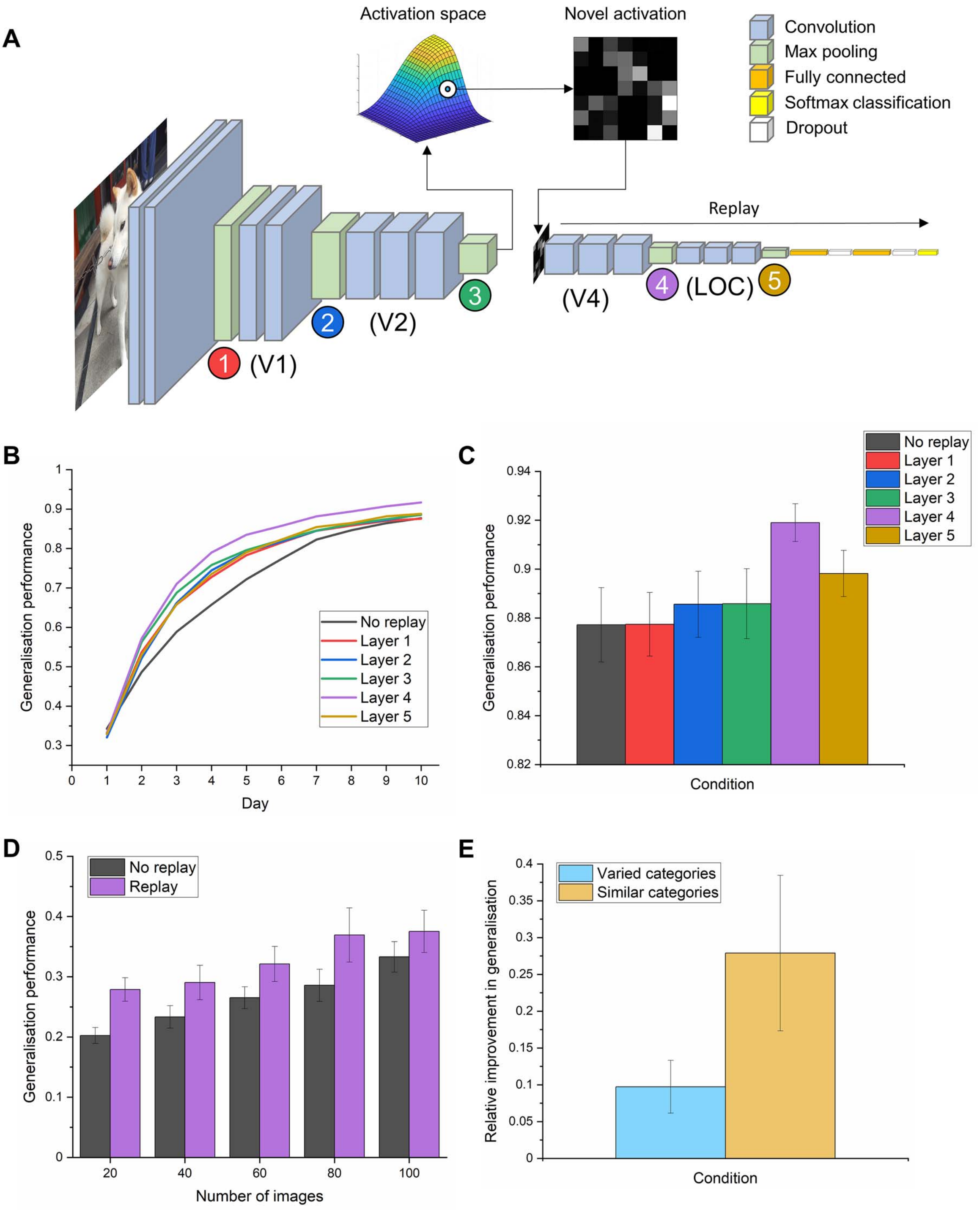

Fig. 1. The effects of generative replay from different layers of a model of the human ventral visual stream on generalization to new exemplars. A) The VGG-16 network attempts to simulate the brain's visual system by looking at photographs and extracting relevant features to help categorize the objects within. We trained this network on 10 new categories of objects it had not seen before. In between learning episodes, to simulate sleep-facilitated consolidation in humans, we implemented offline memory replay as a generative process. In other words, the network "imagined" new examples of a category based on the distribution of features it has learned so far for that object (activation space) and used these representations (novel representation) to consolidate its memory. The network did not create an actual visual stimulus to learn from; rather it recreated the neuronal pattern of activity that it would typically generate from viewing an object from that category. We display here an example of replaying from a midpoint in the network, but all 5 locations where replay was implemented are indicated by the colored circles. The brain regions which have been reported to contain functionally similar representations to different network layers, derived from Güçlü and van Gerven (2015), are listed beneath. B) The effects of memory replay from different layers on the network's ability to generalize to new examples of the 10 categories throughout the course of 10 learning episodes. Plotted values represent the mean accuracies from 10 different models which each learned 10 new and different categories. C) The final recognition accuracies (+/−S.E.M.), averaged across 10 models, on the new set of photographs after 10 epochs of learning. We reveal the location in a model of the ventral stream where replay maximally enhances generalization performance is an advanced layer which bears an approximate functional correspondence to the LOC in humans. The benefits of replay from other locations were less pronounced, with the earliest layer showing the least benefit to generalization. D) The benefits of replay from layer 4 on generalization performance with limited numbers of exemplars E) The effect of generative replay from layer 4 on the generalization performance of classes when learned alongside diverse categories or where all are conceptually similar.

however, it offered several advantages over traditional generative models. We were not limited by a bottleneck symmetrical architecture, and our procedure allowed the model to learn generative samples at multiple levels of representation. Further, our model represented a proper vision model which showed parallels with the functional architecture of the ventral visual stream in the brain, whereas current generative models do not show this correspondence or scale well to such a deep architecture. Finally, our model is specialized for object recognition, with the resulting generated representations shaped by these task pressures.

The number of novel representations created for replay was equivalent to the number of original training images (1,170). To test where in the network replay is most effective, this process was performed at 1 of 5 different network locations, namely the max pooling layers at the end of each block (Fig. 1A). For the first 4 pooling layers, creating a multivariate distribution from such a large number of units was computationally intractable, therefore, activations for each filter in these layers were first downsampled by a factor of 8 for layer 1, by 4 for layers 2 and by 3 and 2 for layer 4. The samples drawn from the resulting distribution were then upsampled back to their original resolution. These lower-resolution samples are also theoretically relevant in that they were created to mimic the schematic nature of mental and dream imagery which takes place during rest and sleep. To replay these samples through the network, the VGG-16 network was temporarily disconnected at the layer where replay was implemented, and a new input layer was attached which matched the dimensions of the replay representations. This truncated network was trained on the replay samples using the same parameters as regular training. We assume that the brain actively chooses to replay each concept learned that day by reactivating the prototypical representations extracted from many experiences and the associated category label together during sleep. After each epoch of replay training, the replay section of the network was reattached to the original base, and training on real images through the whole network resumed. To assess the effects of generative replay on stimuli disambiguation, we took 10 classes from the 100 which were highly similar (plants, see Supplementary Table 2) and trained an additional network on these categories. We then assessed whether replaying similar classes in the same model led to a greater relative increase in class performance from baseline accuracies. We did this by dividing the increase in generalization performance resulting from replay by the original baseline performance. To find out how many exemplars are needed for generative replay to have a beneficial effect on category learning, we trained the same models with 20, 40, 60, 80, and 100 images, again for 10 "days," and replayed an equivalent number of generated representations in each case. To simulate veridical replay, in other words, the replay of each individual experience as it happened, rather than the generation of new samples, we used the activations for each object at that layer in the network during replay periods. These were not downsampled during the process. Given how many examples of a concept we generally encounter, veridical replay of all experience is not a realistic prospect, which is why prior attempts to simulate replay in smaller-scale networks have also avoided this scenario in their approaches (Kemker and Kanan 2017; van de Ven et al. 2020). To additionally demonstrate the improvements that replay affords on each day relative to the previous day, we calculated the performance improvement from day n to day $n+1$, divided by the difference between model performance on day n and 1, which represents the potential room for improvement.

Replay within a reinforcement learning framework

We tested a process through which items that are most beneficial for replay might be selected in the brain. We proposed that such selective replay may involve an interaction between the main concept learning network (VGG-16) and a smaller network which learned through reinforcement which concepts are most beneficial to replay through the main network during offline periods. The neural analog of such a network could be thought of as the hippocampus, as the activity of this structure precedes the widespread reactivation of neural patterns observed during replay (Zhang et al. 2018). This approach is similar to the "teacher–student" meta-learning framework which has been shown to improve performance in deep neural networks (Fan et al. 2018). The side network was a simple regression network with 10 inputs, 1 for each class, and 1 output, which was the predicted value for replaying that class through the main network. Classes were chosen and replayed one at a time, with a batch size of 36. To train the side network, a value of 1 was inputted for the chosen class, with 0s for the others. The predicted reward for the side network was the change in performance of the main network after each replay instance, which was quantified by a change in chi-square; a measure of the difference between the maximum number of possible correct predictions by the main network, versus its actual correct predictions. A positive reward was therefore a reduction in chi-square, which resulted in an increase in the side network's weight for that class. This led to the class being more likely to be chosen in future, as the network's weights were converted into a SoftMax layer from which classes were selected probabilistically for replay. Through this iterative process, the side network learned which classes were more valuable to replay and continually updated its preferences based on the performance of the main network. Reducing the chi-square in this dynamic manner improves the overall network accuracy as it progressively reduces the disparity between the network's classifications and the actual class identities. To generate initial values for the side network, 1 batch of each class was replayed through the main network. The Adam optimizer was used with a learning rate of 0.001 and the objective was to minimize the mean squared error loss. The side network was trained for 50 epochs with each replay batch. The assessment of network improvement was always performed on the validation set, and the reported values are accuracy on the test set, reflecting the ability of the network to generalize to new situations.

Results

Localizing where in the ventral visual stream generative replay is likely to enhance generalization

We first sought to establish where in the visual brain the replay of category knowledge might be most effective in helping to generalize to new experiences, as the functional relevance of replay observed in many different brain regions has yet to be established. To obtain a baseline measure of how the network would perform without replay, the network learned 10 new categories in the absence of offline replay. Next, we implemented generative memory replay. To do this, we captured the "typical" activation of the network for a category and sampled from this gist-like representation to create novel, abstracted representations for replay (Fig. 1A).

We simulated generative replay from different layers in the DCNN, which were equivalent to different brain regions along the ventral stream. Specifically, we trained the network over 10 epochs, mimicking 10 days of learning in humans, and replayed prototypical representations after each training epoch, simulating 10 nights of offline consolidation during sleep. In Fig. 1B, we show how replay affects the ability of the network to generalize to new exemplars of the categories over the course of learning. Replay substantially speeds up the learning process, with replay from layer 4 already reaching the final baseline generalization performance 3 days earlier. Figure 1C shows the final best performing models in each replay condition. A 1-way repeated-measures ANOVA on the final models revealed a difference across conditions ($F_{(5, 45)} = 7.23$, $P < 0.001$), with planned Bonferroni-corrected post hoc comparisons revealing that only replay from layer 4 ($t_{(9)} = -4.31$, $P = 0.002$) was significantly higher than baseline. We performed an additional analysis to confirm that the downsampling of earlier layers did not explain this finding by further downsampling the replay representations in layer 4 by a factor of 7, and generalization performance in this layer was still be significantly higher than baseline (see Supplementary Fig. 2). Therefore, there is a differential benefit of replay throughout the network, where replay in the early layers is of limited benefit, whereas replay in the later layers boosts generalization performance to a greater degree. This predicts that early visual areas in the brain may not store sufficiently complex category-specific representations, curtailing the effectiveness of generated replay representations, whereas areas further along the ventral visual stream, such as the lateral occipital cortex (LOC), might be better positioned to support the generation of novel, prototypical concepts, which accelerates learning in the absence of real experience and helps us to generalize to new situations. We further investigated if generative replay could benefit category learning where few exemplars are available. In Fig. 2D, we show that generative replay from layer 4 could improve generalization when learning and replaying just 20, 40, or 60 exemplars (all *t*-tests below Bonferroni-corrected threshold of $P = 0.01$). We also assessed the effects of replay on class disambiguation in this layer by training a model containing conceptually highly similar classes collated from all of the other models and by comparing the relative increase in generalization performance from the original class accuracies. Figure 2E shows a replay-induced performance increase for conceptually similar items, but this did not reach statistical significance ($t_{(9)} = -2.10$, $P = 0.065$).

Tracking the benefits of replay across learning

In the second experiment, we extended training to 30 days of experience, which were interleaved with nights of offline generative replay, to simulate learning over longer timescales and predict when in learning replay might be more effective (Fig. 2A). Guided by the results of experiment 1, we implemented replay from an advanced layer corresponding to the LOC. A mixed between-within ANOVA showed an interaction between condition and day ($F_{(29,522)} = 5.03$, $P < 0.001$) with planned post hoc Bonferroni-corrected comparisons ($P < 0.00167$) revealing a difference between generative replay and baseline for days 2–6 and 8 (Fig. 2B). Visualizing the network's improvement in performance from day to day relative to the potential room for improvement from the previous day confirmed that the benefits of generative replay were limited to early learning (Fig. 2C). Therefore, offline generative replay might be more effective at improving generalization to new exemplars at the earliest stages of learning. This suggests replay might facilitate rapid generalization, which maximizes performance, given a limited set of experiences with a category.

We were interested to compare generative replay with the unlikely veridical, high-resolution scenario, whereby humans could replay thousands of encounters with individual objects exactly as they were experienced. We termed this "veridical replay" (Fig. 2A), which involved capturing the exact neural patterns associated with each experienced object during learning and replaying these from the same point in the network. A mixed between-within ANOVA did not reveal any difference between generative and veridical replay in terms of generalization performance ($F_{(1,18)} = 0.16$, $P = 0.696$), nor was an interaction effect observed between day and condition ($F_{(29,522)} = 0.29$, $P = 0.999$, Fig. 2B). Therefore, generative replay was comparably effective to veridical replay of experience in consolidating memory despite being entirely imagined from the networks prior experience. This provides tentative support for the hypothesis that

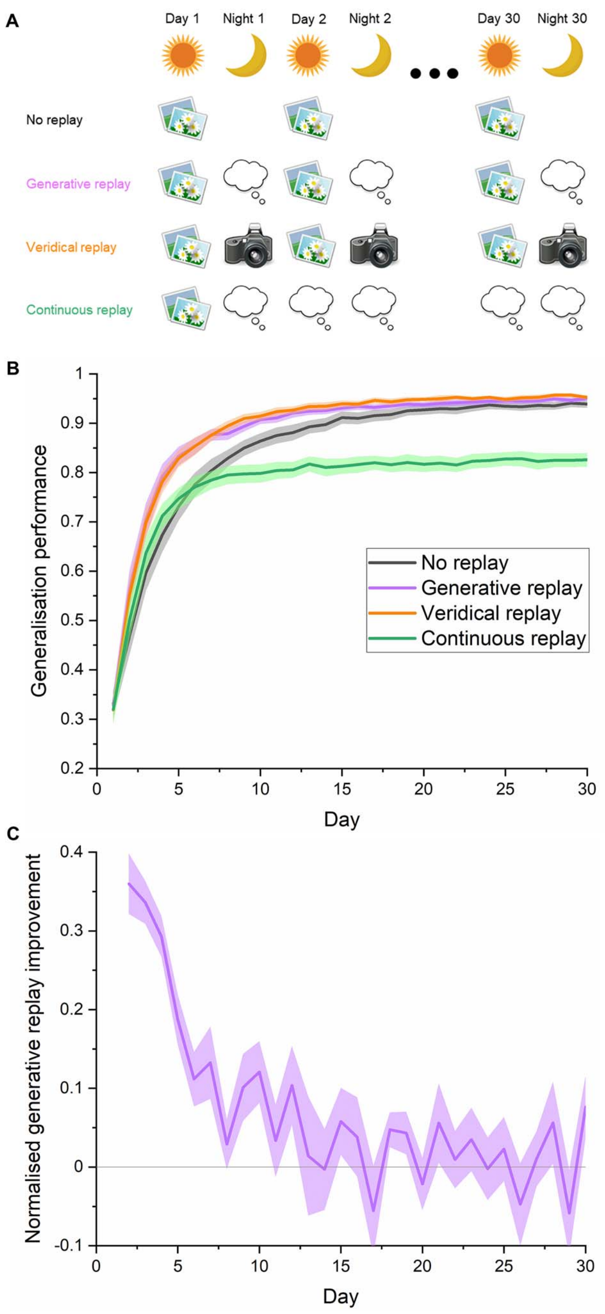

Fig. 2. The facilitatory effects of memory replay across category learning**.** We simulate the long-term consolidation of category memory by extending training to 30 days. A) Schematic showing the different experimental conditions. "No replay" involves the model of the visual system learning the 10 new categories without replay in between episodes. "Generative replay" simulates the brain imagining and replaying novel instances of a category during "night" periods of offline consolidation from a layer bearing some functional approximation to the LOC. "Veridical replay" simulates the hypothetical performance of a human who, each night, replays every single event which has been experienced the preceding day. "Continuous replay" simulates a single day of learning, followed by days and nights of replay, investigating the potential benefit afforded by replay given only brief exposure to a category. For both daytime learning of real images and nighttime consolidation of generated representations, the number of training stimuli was always 1,170 for each class. B) The ability of the network to generalize to new exemplars of a category during each day throughout the learning process. Generalization performance is measured by the proportion (+/−S.E.M.) of correctly recognized test images across 10 models. Generative replay maximally increases performance early in training, suggesting it might be optimal for new learning and recent memory consolidation. Despite being comprised of

generative replay is a putative form of category replay in humans, as it would appear to be vastly more efficient to imagine new concepts from an extracted prototype.

The aforementioned results simulated the benefits of replay under optimal conditions where humans encounter the same categories every day, however, there are instances where exposure will be limited. To what extent can offline replay compensate for this limited learning? We simulated this in our model of the ventral stream by limiting the learning of actual category photographs to 1 day and substituted all subsequent learning experiences with offline replay, which was termed "continuous replay" (Fig. 2A). Despite the absence of further exposure to the actual objects, we found the network could increase its generalization accuracy from 32% to 83% purely by replaying imagined instances of concepts it has partially learned. This result may inform our understanding of the human ability to quickly learn from limited experience. However, a mixed between-within ANOVA revealed a statistically significant interaction effect between day and condition ($F_{(29,522)} = 3.78$, $P < 0.001$), with planned Bonferroni post hoc comparisons revealing a difference between generative replay and continuous replay from day 3 onward (all $P < 0.00167$). Therefore, replayed representations appear to be dynamic in nature, as the prototypes generated from that first experience were not sufficient to train the network to its maximum performance, as is observed when learning and replay are interleaved. This suggests that replayed representations continue to improve as they are informed by ongoing learning, therefore, generative replay in the human brain throughout learning may be envisaged as a constantly evolving "snapshot" of what has been learned so far about that category.

Determining how the brain might select experiences for replay

We proposed that replay may be a learning process in itself, whereby the hippocampus selects replay items and learns through feedback from the neocortex the optimal ones to replay. In our previous simulations, we selected all categories for replay in equal number; however, to simulate the autonomous nature of replay selection in the brain, we supplemented our model of the ventral visual stream with a small reinforcement learning (RL) network, approximating the theoretical role of the hippocampus in deciding what to replay (Fig. 3A). The hippocampus-like model could choose one of the 10 categories to replay and received a reward from the main network for that action based on the improvement in network performance.

We trained our model of the visual system on 10 novel categories, implementing replay during offline periods as before and compared its generalization performance with that of the dual interactive hippocampal-cortical model. In terms of overall accuracy, although generative RL replay appeared to lag briefly behind generative replay at the beginning of training, both approaches performed similarly, with a mixed between-within ANOVA revealing no difference between the 2 conditions in terms of generalization performance ($F_{(1,18)} = 0.15$, $P = 0.704$), nor was an interaction effect observed between day and condition ($F_{(29,522)} = 1.28$, $P = 0.153$, Fig. 3B). Figure 3C plots the difference between the 2 conditions across learning. However, the RL network which simulated the hippocampal replay systematically selected categories which were originally relatively weakly learned more often ($R^2 = 0.24$, $F_{(1,98)} = 31.15$, $P < 0.001$, Fig. 3D), which resulted in their selective improvement ($R^2 = 0.18$, $F_{(1,98)} = 21.15$, $P < 0.001$). However, this came at a cost, with originally well-learned categories being replayed less often and with a drop in their generalization accuracy. We present the idea that such a RL process may underlie the "rebalancing" of experience in the brain and that replay may therefore help to compensate for the fact that some categories are more difficult to learn than others.

Discussion

We simulated the consolidation of category knowledge in a large-scale neural network model which approximates the functional aspects of the human ventral visual system by replaying prototypical representations thought to be formed and initiated by the hippocampus. The notion that replay of visual experiences might be generative in nature has been suggested by limited-capacity models which have been trained on low-resolution photographic images (van de Ven et al. 2020). However, our results using a model of the visual brain, whose representations has compared favorably with actual brain data, represent more persuasive evidence that humans are unlikely to replay experiences verbatim during rest and sleep to improve category knowledge and might be more likely to replay novel, imagined instances instead. In addition, the large number (117,000) of high-resolution complex naturalistic images we used for training in this experiment more closely reflected real-world learning and facilitated the extraction of gist-like features. While empirical evidence exists that humans replay novel sequences of stimuli (Liu et al. 2019), our work suggests that the brain might go further and use learned features of objects to construct entirely fictive experiences to replay. We speculate that this creative process is particularly important for the consolidation of category knowledge as opposed to the replay of episodic memory (Deuker et al. 2013; Schapiro et al. 2018; Zhang et al. 2018) because of the requirement to abstract prototypical features and use these to generalize to new examples of a category. We propose that generative replay confers additional advantages, such as constituting less of a burden on memory resources, as not all experiences need to be remembered. Further, our replay representations were highly effective in consolidating category knowledge despite being downsampled, and these compressed, low-resolution samples would reduce storage requirements further. Perhaps the simulation that most favorably supported the hypothesis that category replay in the brain likely adopts this compressed, prototypical format is that it aided generalization to a similar degree as the exact veridical replay of experience in boosting generalization performance. Therefore, the main advantage to generative replay over veridical replay is that it represents a feasible, efficient solution to memory consolidation without compromising effectiveness. In addition, generative replay can add to events which have been experienced. Our findings therefore encourage a reconceptualization of the nature of consolidation-related replay in humans that it is not only generative but also low resolution or "blurry," as is the case with internally generated imagery in humans (Giusberti et al. 1992; Lee et al. 2012). In fact, the kind of replay we propose here may be the driving force behind the transformation of memory into a more schematic, generalized form which preserves regularities across experiences while allowing unique elements of experience to fade (Love and Medin 1998; Winocur and Moscovitch 2011; Sweegers and Talamini 2014). The challenge for future empirical studies in humans to confirm our hypothesis will be to decode prototypical replay representations during rest and sleep. In addition, future modeling and empirical work should address the sequential nature of learning and replay, as life experience does not consist of still snapshots of experience, such as those used in these experiments. Prior modeling work has shown that a video game-playing agent can improve its performance by learning inside its own generated environment (Ha and Schmidhuber 2018), which is more akin to an unfolding dream during sleep and may provide inspiration for modeling the generative replay of video-like events to support category learning.

Simulating replay in a human-like network also allowed us to answer a question not currently tractable in neuroimaging studies: Where in the visual stream is replay functionally relevant to consolidation? In a prior simulation of replay in a neural network, van de Ven et al. (2020) demonstrated generative replay could attenuate forgetting when performed after the final convolutional layer, but its effectiveness was not compared to earlier layers, and the network employed, consisting of 5 convolutional layers, had not been compared with the human visual system. Deeper networks, such as the one used here, consisting of 23 layers in total, organized

internally generated fictive experiences, generative replay was comparably effective to veridical replay throughout the learning process, positing it as an attractive, efficient, and more realistic solution to memory consolidation which does not involve remembering all experiences. Continuous replay after just 1 day of learning substantially improved generalization performance but never reached the accuracy levels of networks which engaged in further learning. C) The improvement in performance that generative replay affords on each day relative to the possible improvements from the previous day.

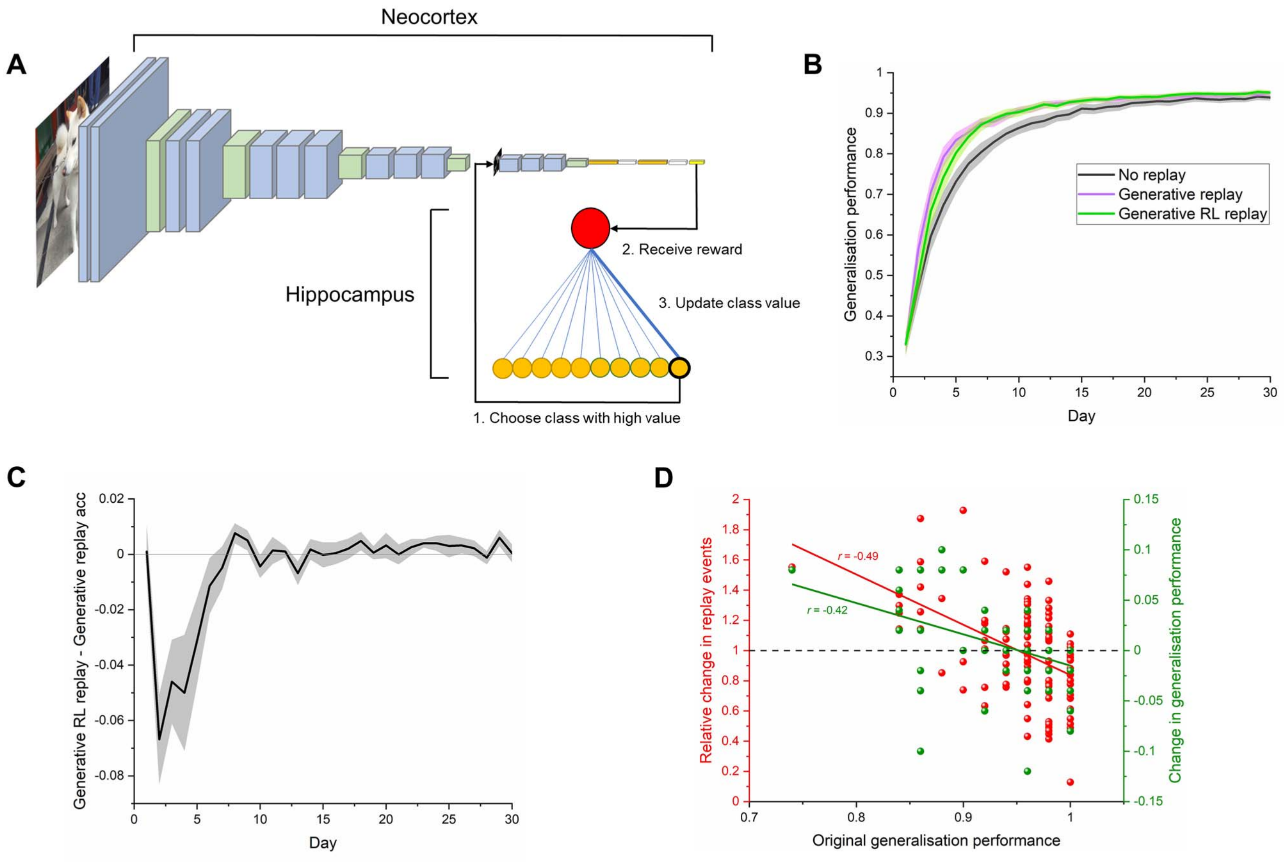

Fig. 3. Replay as a RL process simulates the brain's tendency to consolidate weaker knowledge. A) Replay in a model which approximates the visual system is controlled by a RL network which aims to assume the role of the hippocampus. The RL network selects 1 of 10 categories to replay through the visual system and receives a reward based on the improved performance, learning through trial and error which categories to replay. B) Overall generalization performance on new category exemplars was similar for both generative replay and generative replay controlled by a RL network. Generalization performance represents mean accuracy (+/−S.E.M) on test images across 10 models which each learned 10 new categories. C) The difference between generative replay and generative RL replay performance for each day. D) The RL network learns to replay categories which were originally more difficult for the model of the visual system and improves their accuracy. This effectively "rebalanced" memory such that category knowledge was more evenly distributed and offers a candidate mechanism as to how the brain chooses weakly learned information for replay. Plotted values represent the 100 categories across 10 models. A proportion of the generalization performance values are overlapping.

into 5 blocks of convolutional layers, not only extract useful category features from naturalistic images but representations in network layers have also demonstrated a degree of representational correspondence with specific brain regions along the ventral visual stream (Khaligh-Razavi and Kriegeskorte 2014; Güçlü and van Gerven 2015; Devereux et al. 2018), albeit not capturing all observable variance (Xu and Vaziri-Pashkam 2021). In keeping with our observation that low-resolution, coarse, schematic replay was effective in helping the network to generalize, we found the most effective location for replay to be in the most advanced layers of the network, layers which are less granular in their representations. This region shares some functional similarities with the LOC in humans, a region which represents more complex, high-level features (Güçlü and van Gerven 2015). By contrast, generative replay from the earliest layers corresponding to early visual cortex was less effective. These layers are sensitive to low-level visual features, such as contrast, edges, and color, therefore generating samples from these layers will yield rudimentary-level category-specific information, which is of limited utility for replay and generalization. High-level representations on the other hand, may contain more unique combinations and abstractions of these lower-level features. We also found that replay from the penultimate layer was more effective than the final layer, suggesting the optimal replay location represents a balance between the presence of sufficiently complex category information and the number of downstream neuronal weights available to be updated based on replaying these features. These findings may encourage a reevaluation of the functional relevance of replay in early visual cortices in both animals and humans and generate specific hypotheses for potential perturbation studies to investigate the effects of disruptive stimulation at different stages of the ventral stream during offline consolidation.

Our simulations also revealed a phenomenon never before tested in humans that the effectiveness of replay depends on the stage of learning. We acquire factual information about the world sporadically over time across contexts, for example, we may encounter a new species at a zoo 1 day and subsequently see the same animal on a wildlife documentary, and so on. Ultimately, the consolidation of semantic information in the neocortex can take up to years to complete (Manns et al. 2003). However, our simulations suggest that replay may be most beneficial during the initial encounters with a novel category, when we are still working out its identifiable features and have not yet learned to generalize perfectly to unseen instances. It is therefore possible that humans replay a category less and less with increasing familiarity, and there is some support for this idea in the animal literature (Giri et al. 2019). We speculate that, if this is the case, the enhanced effectiveness for recent memories may have an adaptive function, allowing us to generalize quickly with limited information. In fact, our simulations showed that after a single learning episode, replay can compensate substantially for an absence of subsequent experience. Our results provide novel hypotheses for human experiments, testing for an interaction between the stage of category learning and the extent of replay. The fact that replay early in the learning process was more effective provides further support for our proposal that vague, imprecise replay events are useful for generalization, as the networks imaginary representations at that stage would be an imperfect approximation of the category in question. We acknowledge there may be a "ceiling effect," whereby later in training there is no further room for improvement; however, we would posit that over the human lifespan, we are operating in the nonconverged portion of the learning curve that we display here.

Our results also represent the first mechanistic account of how the brain selects weakly learned information for replay and consolidation (Kuriyama et al. 2004; Drosopoulos et al. 2007; McDevitt et al. 2015; Schapiro et al. 2018). The hippocampus triggers replay events in the neocortex (Zhang et al. 2018), with a loop of information back and forth between the two brain areas (Rothschild et al. 2017), although the content of this neural dialog is not known. Our simulations suggest that the hippocampus may learn the optimal categories to replay based on feedback from the neocortex. Our results showed that such a process resulted in the "rebalancing" of experience in an artificial neural network, where generalization performance was improved for weakly learned items and attenuated for items which were strongly learned. A reorganization of knowledge of this kind has been observed in electrophysiological investigations in rodents, where the neural representations of novel environments are strengthened through reactivation at the peak of the theta cycle, while those corresponding to familiar environments are weakened through replay during the trough (Poe et al. 2000). This more even distribution of knowledge could be adaptive in both ensuring adequate recognition performance across all categories and forming a more general foundation on top of which future conceptual knowledge can be built. There have been recent theoretical accounts and empirical demonstrations of how items get selected for replay within a reinforcement learning framework. These include the prioritization of events for replay which were most surprising during learning (Momennejad et al. 2018), and the replay of events that are more likely to be encountered in future and which lead to the highest reward (Mattar and Daw 2018; Liu et al. 2021). However, these accounts do not explain why even in the absence of such prediction errors, or without knowing the likelihood of future events, knowledge which has been weakly learned during waking periods is consistently targeted for replay and consolidation during sleep (Kuriyama et al. 2004; Drosopoulos et al. 2007; McDevitt et al. 2015; Schapiro et al. 2018). Our interactive networks suggest that offline RL could account for the selection of weakly learned knowledge during the replay process itself, and future experiments could assess whether our models choose the same categories for replay as humans when trained on the same stimuli.

Conclusion

In summary, our simulations provide supportive evidence that category replay in humans is a generative process and make the prediction that it is functionally relevant at advanced stages of the ventral stream. We have generated hypotheses about when during learning replay is likely to be effective and offer a novel account of replay as a learning process in and of itself between the hippocampus and neocortex. We hope these findings encourage a closer dialog between theoretical models and empirical experiments. These findings also add credence to the emerging perspective that deep learning networks are powerful tools which are becoming increasingly well positioned to resolve challenging neuroscientific questions (Richards et al. 2019).

Acknowledgements

We thank the Love Lab for helpful discussions on this project. This work was supported by the Royal Society (grant number 183029) and the Wellcome Trust (grant number WT106931MA) to B.C.L. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary material is available at *Cerebral Cortex Journal* online.

Conflict of interest statement. None declared.

Authors' contributions

D.N.B. and B.C.L. were in charge of conceptualization, methodology, and writing—review and editing. D.N.B. was in charge of software, data curation, investigation, formal analysis, visualization, and writing—original draft preparation. B.C.L. was in charge of resources, funding acquisition, and supervision.

Data and code availability

The data underlying this article are available in Figshare at https://figshare.com, and can be accessed with https://doi.org/10.6084/m9.figshare.14208470.

References

Bowman CR, Iwashita T, Zeithamova D. Tracking prototype and exemplar representations in the brain across learning. *elife*. 2020:9:e59360.

de Voogd LD, Fernández G, Hermans EJ. Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. *NeuroImage*. 2016:134:563–572.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE: New York. 2009. p. 248–255.

Deuker L, Olligs J, Fell J, Kranz TA, Mormann F, Montag C, Reuter M, Elger CE, Axmacher N. Memory consolidation by replay of stimulus-specific neural activity. *J Neurosci*. 2013:33: 19373–19383.

Devereux BJ, Clarke A, Tyler LK. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci Rep*. 2018:8:10636.

Drosopoulos S, Windau E, Wagner U, Born J. Sleep enforces the temporal order in memory. *PLoS One*. 2007:2:e376.

Dudai Y. The restless engram: consolidations never end. *Annu Rev Neurosci*. 2012:35:227–247.

Eichenlaub J-B, Jarosiewicz B, Saab J, Franco B, Kelemen J, Halgren E, Hochberg LR, Cash SS. Replay of learned neural firing sequences during rest in human motor cortex. *Cell Rep*. 2020:31:107581.

Ekman M, Kok P, de Lange FP. Time-compressed preplay of anticipated events in human primary visual cortex. *Nat Commun*. 2017:8:15276.

Fan Y, Tian F, Qin T, Li X-Y, Liu T-Y. Learning to teach. 2018. arXiv preprint arXiv:180503643

Fenn KM, Nusbaum HC, Margoliash D. Consolidation during sleep of perceptual learning of spoken language. *Nature*. 2003:425: 614–616.

French RM. Catastrophic forgetting in connectionist networks. *Trends Cogn Sci*. 1999:3:128–135.

Friedrich M, Wilhelm I, Born J, Friederici AD. Generalization of word meanings during infant sleep. *Nat Commun*. 2015:6:6004.

Girardeau G, Inema I, Buzsáki G. Reactivations of emotional memory in the hippocampus–amygdala system during sleep. *Nat Neurosci*. 2017:20:1634.

Giri B, Miyawaki H, Mizuseki K, Cheng S, Diba K. Hippocampal reactivation extends for several hours following novel experience. *J Neurosci*. 2019:39:866–875.

Giusberti F, Cornoldi C, De Beni R, Massironi M. Differences in vividness ratings of perceived and imagined patterns. *Br J Psychol*. 1992:83:533–547.

Gómez RL, Bootzin RR, Nadel L. Naps promote abstraction in language-learning infants. *Psychol Sci*. 2006:17:670–674.

Graveline YM, Wamsley EJ. The impact of sleep on novel concept learning. *Neurobiol Learn Mem*. 2017:141:19–26.

Grossberg S. Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw*. 2013:37:1–47.

Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J Neurosci*. 2015:35:10005–10014.

Gupta AS, van der Meer MA, Touretzky DS, Redish AD. Hippocampal replay is not a simple function of experience. *Neuron*. 2010:65: 695–705.

Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution. *Adv Neural Inf Process Syst*. 2018:31:2450–2462.

Hassabis D, Kumaran D, Vann SD, Maguire EA. Patients with hippocampal amnesia cannot imagine new experiences. *Proc Natl Acad Sci U S A*. 2007:104:1726–1731.

Hayes TL, Krishnan GP, Bazhenov M, Siegelmann HT, Sejnowski TJ, Kanan C. Replay in deep learning: current approaches and missing biological elements. 2021. arXiv preprint arXiv:2104.04132

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *Proc IEEE Int Conf Comput Vis*. 2015:1026–1034.

Hermans EJ, Kanen JW, Tambini A, Fernández G, Davachi L, Phelps EA. Persistence of amygdala–hippocampal connectivity and multi-voxel correlation structures during awake rest after fear learning predicts long-term expression of fear. *Cereb Cortex*. 2017:27:3028–3041.

Horváth K, Liu S, Plunkett K. A daytime nap facilitates generalization of word meanings in young toddlers. *Sleep*. 2016:39:203–207.

Jafarpour A, Fuentemilla L, Horner AJ, Penny W, Duzel E. Replay of very early encoding representations during recollection. *J Neurosci*. 2014:34:242–248.

Ji D, Wilson MA. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat Neurosci*. 2007:10:100–107.

Kemker R, Kanan C. Fearnet: brain-inspired model for incremental learning. 2017. arXiv preprint arXiv:171110563.

Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*. 2014:10:e1003915.

Kingma DP, Ba J. Adam: a method for stochastic optimization; 2014. arXiv preprint arXiv:14126980

Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci U S A*. 2017:114:3521–3526.

Kuriyama K, Stickgold R, Walker MP. Sleep-dependent learning and motor-skill complexity. *Learn Mem*. 2004:11:705–713.

Lau H, Alger SE, Fishbein W. Relational memory: a daytime nap facilitates the abstraction of general concepts. *PLoS One*. 2011:6:e27139.

Lee SH, Kravitz DJ, Baker CI. Disentangling visual imagery and perception of real-world objects. *NeuroImage*. 2012:59:4064–4073.

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ. Human replay spontaneously reorganizes experience. *Cell*. 2019:178:640–652.

Liu Y, Mattar MG, Behrens TEJ, Daw ND, Dolan RJ. Experience replay is associated with efficient nonlocal learning. *Science*. 2021:372:6544.

Love BC, Medin DL. SUSTAIN: a model of human category learning. Aaai/iaai. 1998:671–676.

Lutz ND, Diekelmann S, Hinse-Stern P, Born J, Rauss K. Sleep supports the slow abstraction of gist from visual perceptual memories. *Sci Rep*. 2017:7:42950.

Manns JR, Hopkins RO, Squire LR. Semantic memory and the human hippocampus. *Neuron*. 2003:38:127–133.

Mattar MG, Daw ND. Prioritized memory access explains planning and hippocampal replay. *Nat Neurosci*. 2018:21:1609–1617.

McDevitt EA, Duggan KA, Mednick SC. REM sleep rescues learning from interference. *Learn Mem*. 2015:122:51–62.

McGaugh JL. Memory—a century of consolidation. *Science*. 2000:287: 248–251.

Michelmann S, Staresina BP, Bowman H, Hanslmayr S. Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nat Hum Behav*. 2019:3:143–154.

Momennejad I, Otto AR, Daw ND, Norman KA. Offline replay supports planning in human reinforcement learning. *elife*. 2018:7:e32548.

Peyrache A, Khamassi M, Benchenane K, Wiener SI, Battaglia FP. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat Neurosci*. 2009:12:919–926.

Poe GR, Nitz DA, McNaughton BL, Barnes CA. Experience-dependent phase-reversal of hippocampal neuron firing during REM sleep. *Brain Res*. 2000:855:176–180.

Remy P. 2020. *Keract: a library for visualizing activations and gradients*. GitHub repository. https://github.com/philipperemy/keract.

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S, et al. A deep learning framework for neuroscience. *Nat Neurosci*. 2019:22:1761–1770.

Robins A. Catastrophic forgetting, rehearsal and Pseudorehearsal. *Connect Sci*. 1995:7:123–146.

Rothschild G, Eban E, Frank LM. A cortical-hippocampal-cortical loop of information processing during memory consolidation. *Nat Neurosci*. 2017:20:251–259.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015:115:211–252.

Schapiro AC, McDevitt EA, Chen L, Norman KA, Mednick SC, Rogers TT. Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Sci Rep*. 2017:7:14869.

Schapiro AC, McDevitt EA, Rogers TT, Mednick SC, Norman KA. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat Commun*. 2018:9:3920.

Schönauer M, Alizadeh S, Jamalabadi H, Abraham A, Pawlizki A, Gais S. Decoding material-specific memory reprocessing during sleep in humans. *Nat Commun*. 2017:8:15404.

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Schmidt K, *et al*. Brain-score: Which artificial neural network for object recognition is most brain-like?. 2018. bioRxiv.407007.

Schwartenbeck P, Baram A, Liu Y, Mark S, Muller T, Dolan R, Botvinick M, Kurth-Nelson Z, Behrens T. Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. 2021. bioRxiv.2021.2006.2006.447249.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:14091556.

Squire LR, Alvarez P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr Opin Neurobiol*. 1995:5: 169–177.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014:15:1929–1958.

Staresina BP, Alink A, Kriegeskorte N, Henson RN. Awake reactivation predicts memory in humans. *Proc Natl Acad Sci U S A*. 2013:110: 21159–21164.

Sweegers CCG, Talamini LM. Generalization from episodic memories across time: a route for semantic knowledge acquisition. *Cortex*. 2014:59:49–61.

Tambini A, Davachi L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc Natl Acad Sci U S A*. 2013:110:19591–19596.

Tambini A, Ketz N, Davachi L. Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron*. 2010:65:280–290.

van de Ven GM, Siegelmann HT, Tolias AS. Brain-inspired replay for continual learning with artificial neural networks. *Nat Commun*. 2020:11:4069.

Wilson MA, McNaughton BL. Reactivation of hippocampal ensemble memories during sleep. *Science*. 1994:265:676–679.

Wimmer GE, Liu Y, Vehar N, Behrens TEJ, Dolan RJ. Episodic memory retrieval success is associated with rapid replay of episode content. *Nat Neurosci*. 2020:23:1025–1033.

Winocur G, Moscovitch M. Memory transformation and systems consolidation. *J Int Neuropsychol Soc*. 2011:17:766–780.

Wittkuhn L, Schuck NW. Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nat Commun*. 2021:12:1795.

Xu Y, Vaziri-Pashkam M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat Commun*. 2021:12:2065.

Zhang H, Fell J, Axmacher N. Electrophysiological mechanisms of human memory consolidation. *Nat Commun*. 2018:9:4103.