A neural network account of memory replay and knowledge consolidation

Running title: Category replay in deep neural networks

Daniel N. Barry[1] and Bradley C. Love[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: daniel.barry@ucl.ac.uk, ORCID ID: 0000-0002-2474-5651

Abstract

Replay can consolidate memories through offline neural reactivation related to past experiences. Category knowledge is learned across multiple experiences, and its subsequent generalisation is promoted by consolidation and replay during rest and sleep. However, aspects of replay are difficult to determine from neuroimaging studies alone. Here, we provided insights into category knowledge replay by simulating these processes in a neural network which assumed the roles of the human ventral visual stream and hippocampus. Generative replay, akin to imagining new instances of a category, facilitated generalisation 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 equivalent 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; a reinforcement learning process in which categories were replayed according to their contribution to network performance. This reconceptualises consolidation-related replay as an active rather than passive process.

Keywords: consolidation, learning, memory, network, replay

1. Introduction

Memory replay refers to the reactivation of experience-dependent neural activity during resting periods. First observed in rodent hippocampal cells during sleep (Wilson & McNaughton, 1994), the phenomenon has since been detected in humans during rest (Hermans et al., 2017; Liu, Dolan, Kurth-Nelson, & Behrens, 2019; Schapiro, McDevitt, Rogers, Mednick, & Norman, 2018; Tambini & Davachi, 2013; Wittkuhn & Schuck, 2021), and sleep (Schönauer et al., 2017; Zhang, Fell, & Axmacher, 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. This memory consolidation-related replay 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, Fuentemilla, Horner, Penny, & Duzel, 2014; Michelmann, Staresina, Bowman, & Hanslmayr, 2019; Wimmer, Liu, Vehar, Behrens, & Dolan, 2020), visual understanding (Schwartenbeck et al., 2021), decision making (Liu, Mattar, Behrens, Daw, & Dolan, 2021), planning (Momennejad, Otto, Daw, & Norman, 2018) and prediction (Ekman, Kok, & de Lange, 2017). In this paper we challenge the notion of offline consolidation-related replay as a passive, memory-preserving process, and propose it is much more dynamic in nature. 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.

Neural replay which supports memory consolidation during rest and sleep is generally 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 generalise to new experiences rather than recognise past instances, a phenomenon observed after sleep in infants (Friedrich, Wilhelm, Born, & Friederici, 2015; Horváth, Liu, & Plunkett, 2016). 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, Kumaran, Vann, & Maguire, 2007), we consider this 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 re-ordered based on previously learned rules (Liu et al., 2019). This is reminiscent of "pre-play" 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, van der Meer, Touretzky, & Redish, 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 it's flexibility may not just apply to the reorganisation of learned sequences, but the creation of entirely new instances of a category. While decoding the re-ordering 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 modelled the replay of generated image stimuli (van de Ven, Siegelmann, & Tolias, 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 five 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

mapped directly on to specific visual brain regions indicating a close correspondence in functional architecture (Devereux, Clarke, & Tyler, 2018; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014), and whose object recognition performance compares favourably with humans (He, Zhang, Ren, & Sun, 2015). Further, the networks employed by van de Ven et al. (2020) had limited visual experience, having been pre-trained on just 10 categories of objects. In contrast, an adult human brain will harbour 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 pre-existing "lifetime" of knowledge about a vast 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 x 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 reflect 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 & 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 (French, 1999; Robins, 1995). As biological agents do not suffer from this issue, the findings of these studies offer little insight into human brain and behaviour.

In this study, we investigated whether offline generative replay of novel concepts facilitated subsequent generalisation to new experiences using models which reflect the human brain and 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 organised into five blocks of convolutional layers and boasts a high "brain-score" indicating it extracts representations in a similar manner to the brain and performs comparably to humans in a categorisation task (Schrimpf et al., 2018). The network had prior experience of learning 1000 diverse categories of objects from over a million high-resolution complex naturalistic images, a process which

rivals a lifetime of human visual experience and which yields within the model the equivalent of a mature fully-functioning visual system. We tasked the model with learning 10 novel categories it had not seen before, using similarly high-resolution naturalistic images with an average resolution of around 400 x 350 pixels (Deng et al., 2009), representing an approximate 140-fold increase in visual details from stimuli used in prior work. This is equivalent to a human coming across 10 new categories and using their lifelong experience in processing visual information to extrapolate the relevant identifying features. After learning periods, we then simulated generative replay in the network, akin to human consolidation during sleep, and monitored the network's performance when it "woke up" the next day, to ascertain if such a process could explain the overnight improvements in generalisation observed in humans.

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 towards generalisation. Replay has been observed throughout the brain, early in the ventral visual stream (Deuker et al., 2013; Ji & Wilson, 2007; Wittkuhn & Schuck, 2021), in the ventral temporal cortex (de Voogd, Fernández, & Hermans, 2016; Tambini, Ketz, & Davachi, 2010), the medial temporal lobe (Schapiro et al., 2018; Staresina, Alink, Kriegeskorte, & Henson, 2013) the amygdala, (Girardeau, Inema, & Buzsáki, 2017; Hermans et al., 2017), motor cortex (Eichenlaub et al., 2020) and prefrontal cortex (Peyrache, Khamassi, Benchenane, Wiener, & Battaglia, 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 specified functional correspondence with the human visual brain, and because replay from other locations within the network was not also implemented for comparison, it is difficult to draw conclusions from these results regarding effective replay locations in the human brain. In the current study, because we

153 simulated replay in a neural network which bears a close correspondence with the human ventral
154 visual stream, we could compare the effectiveness of replay from different layers with a known
155 representational correspondence to specific regions in the brain. In doing so, we aimed to find out the
156 effective cortical targets of offline memory consolidation in humans.

157 Another open question regarding human replay is the duration of its involvement throughout
158 the learning of novel concepts. It can take humans years to learn and consolidate semantic or
159 conceptual knowledge (Manns, Hopkins, & Squire, 2003), but neuroimaging studies of replay are
160 limited to a time-span of a day or two, therefore it is still not known how long replay contributes to
161 this process. Humans are thought to "reconsolidate" information every time it is retrieved (Dudai,
162 2012), suggesting replay might play a continual role in the lifespan of memory. However recordings in
163 rodents have shown that replay diminishes with repeated exposure to an environment over multiple
164 days (Giri, Miyawaki, Mizuseki, Cheng, & Diba, 2019), suggesting the brain may only replay recently
165 learned, vulnerable information. Answering this question in humans remains a challenge because of
166 the impracticalities of tracking replay events for extended periods. Simulation in a human-like neural
167 network represents a feasible alternative to determine the relative contribution of replay to
168 consolidation over long time-periods, an approach which has not been attempted to date. Here, we
169 interleaved daily learning with nights of offline replay in a brain-like neural network to understand at
170 what stage in learning replay may be most effective in humans.

171 An additional poorly understood principle of replay which we investigated in this study is why
172 consolidation tends to selectively benefit weakly-learned over well-learned information (Drosopoulos,
173 Windau, Wagner, & Born, 2007; Kuriyama, Stickgold, & Walker, 2004; McDevitt, Duggan, & Mednick,
174 2015; Schapiro et al., 2018). Here, we modelled a candidate mechanism for how this occurs in the
175 brain, by adding an auxiliary model (akin to the hippocampus) to the neocortical model, which could
176 autonomously learn the best consolidation strategy, determining what to replay and when.

177 In addressing these outstanding questions regarding replay in the brain, we made a number
178 of predictions. Because earlier brain regions are thought to extract equivalent basic features from all

179 categories, we predicted replay of experience would be more effective in promoting learning at
180 advanced stages of the network. We hypothesised the replay of "imagined" prototypical replay events
181 would be as effective as veridical replay in helping us to generalise to new, unseen experiences, thus
182 supporting our conceptualization of replay as a creative process. We predicted that the benefits of
183 replay may be confined to early in the learning curve when novel category knowledge is being
184 acquired. Finally, we hypothesised that a dynamic interaction between a hippocampal and neocortical
185 model would result in the prioritisation of weakly-learned items, in line with behavioural studies of
186 memory consolidation.
187

188 ## 2. Materials and Methods

189 ### 2.1 Neural network

190 To simulate the learning of novel concepts in the brain, and test a number of hypotheses regarding
191 replay, we trained a DCNN on 10 new categories of images. The neural network was VGG-16
192 (Simonyan & Zisserman, 2014). Emulating the extent of real-world learning in humans, this network is
193 trained on a vast dataset of 1.3 million high-resolution complex naturalistic photographs known as the
194 ImageNet database (Deng et al., 2009), which contains recognisable objects from 1000 categories in
195 different contexts much like what humans encounter on a daily basis. The network learns to associate
196 the visual features of an object with its category label, until it can recognise examples of that object
197 which it has never seen before, reflecting the human ability to generalise prior knowledge to new
198 situations. The network takes a photograph's pixels as input, and sequentially transforms this input
199 into more abstract features, similar to the operation of the human ventral visual stream (Güçlü & van
200 Gerven, 2015). It learns to perform these transformations by adjusting 138,357,544 connection
201 weights across many layers. Its convolutional architecture reduces the number of possible training
202 weights by searching for informative features in any area of the photographs.

203 This network which has been previously trained on 1000 categories can be thought of as
204 equivalent to a fully functional visual system. This visual system allows humans to rapidly learn new

categories because it facilitates the extraction of useful features to support learning. Similarly, the VGG-16 network can learn novel categories which it has not learned before, based on its prior experience in interpreting visual input. In these experiments, we task the VGG-16 network with learning 10 new categories of images. To do this, we retained take the pre-trained "base" of this network, which consisted of 19 layers, organised into five convolutional blocks. Within each block there were convolutional layers and a pooling layer, with nonlinear activation functions. To this base, we attached two fully connected layers, each followed by a "dropout" layer, which randomly zeroed out 50% of units to prevent overfitting to the training set (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). At the end of the network a SoftMax layer was attached, which predicted which of 10 classes an image belonged to. To facilitate the learning of 10 new classes, the weights of layers attached to the pre-trained base were randomly initialised. All model parameters were free to be trained. In total, 10 new models were trained, each learning 10 new and different classes.

2.2 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 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.

2.3 Baseline training

We first trained a model without implementing replay, to serve as a baseline measure of network performance, and compare 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 equivalent to a human viewing an object at different

231 locations, or from different angles, and facilitates the extraction of useful features rather than rote
232 memorisation of experience. Augmentation could include up to 20-degree rotation, 20% vertical or
233 horizontal shifting, 20% zoom, and horizontal flipping. Any blank portions of the image following
234 augmentation were filled with a reflection of the existing image. Images were then pre-processed in
235 accordance with Simonyan and Zisserman (2014). Depending on the experiment, the network was
236 trained for 10 or 30 epochs. We used the Adam optimiser (Kingma & Ba, 2014) with a learning rate of
237 0.0003. A small learning rate was chosen to reflect the fact that learning new categories in an adult
238 human reflects a "fine-tuning" of an already highly-trained visual system. The training batch size was
239 set to 36. The training objective was to minimise the categorical cross-entropy loss over the 10 classes.
240 Training parameters were optimised based on validation set performance. We report the model's
241 performance metrics from the test set only. This is a collection of novel images from each category
242 which the network does not learn nor is it tuned on, therefore reflecting the model's ability to
243 generalise to new stimuli after training, and is thus termed "generalisation performance" in the
244 figures. Training was performed using TensorFlow version 2.2.
245

246 **2.4 Replay**

247 Replay was conducted between training epochs, to simulate "days" of learning and "nights" of offline
248 consolidation. We conceptualised replay representations as generative, in other words they
249 represented a prototype of that category never seen before, from a particular point in the network.
250 This represents an alternative to storing every experience in our heads, in that we could replay
251 important knowledge about the world without remembering everything. To generate these
252 representations, the network activations induced by the training images from the preceding epoch
253 were extracted from a particular layer in the network using the Keract toolbox (Remy, 2020). For each
254 class separately, a multivariate distribution of activity was created from these activations,
255 representing the unique relationship between units of the layer which were observed for that specific
256 class. We then sampled randomly from this distribution, creating novel activation patterns for that

class at that point in the network (Figure 1). 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 novel in nature. This would be equivalent in the brain to an approximate pattern of neural activity which is representative of that category at a particular stage in the ventral visual stream. These prototypical concepts would be likely generated from more high-level regions such as the hippocampus and prefrontal cortex (Bowman, Iwashita, & Zeithamova, 2020; Hassabis, Kumaran, Vann, et al., 2007).

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 one of five different network locations, namely the max pooling layers at the end of each block (Figure 1). For the first four 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 down-sampled by a factor of eight for layer one, by four for layers two and three and two for layer four. The samples drawn from the resulting distribution were then up-sampled back to their original resolution. These lower-resolution samples are also theoretically relevant, in that they are more akin to 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. 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 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 down-sampled 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 & Kanan, 2017; van de Ven et al., 2020).

2.5 Replay within a reinforcement learning framework

We tested a process through which items which are most beneficial for replay may 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 analogue 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, Tian, Qin, Li, & Liu, 2018). The side network was a simple regression network with 10 inputs, one for each class, and one 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 zeros 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 contrast of 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, one batch

of each class was replayed through the main network. The Adam optimiser was used with a learning rate of 0.001 and the objective was to minimise 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 generalise to new situations.

3. Results

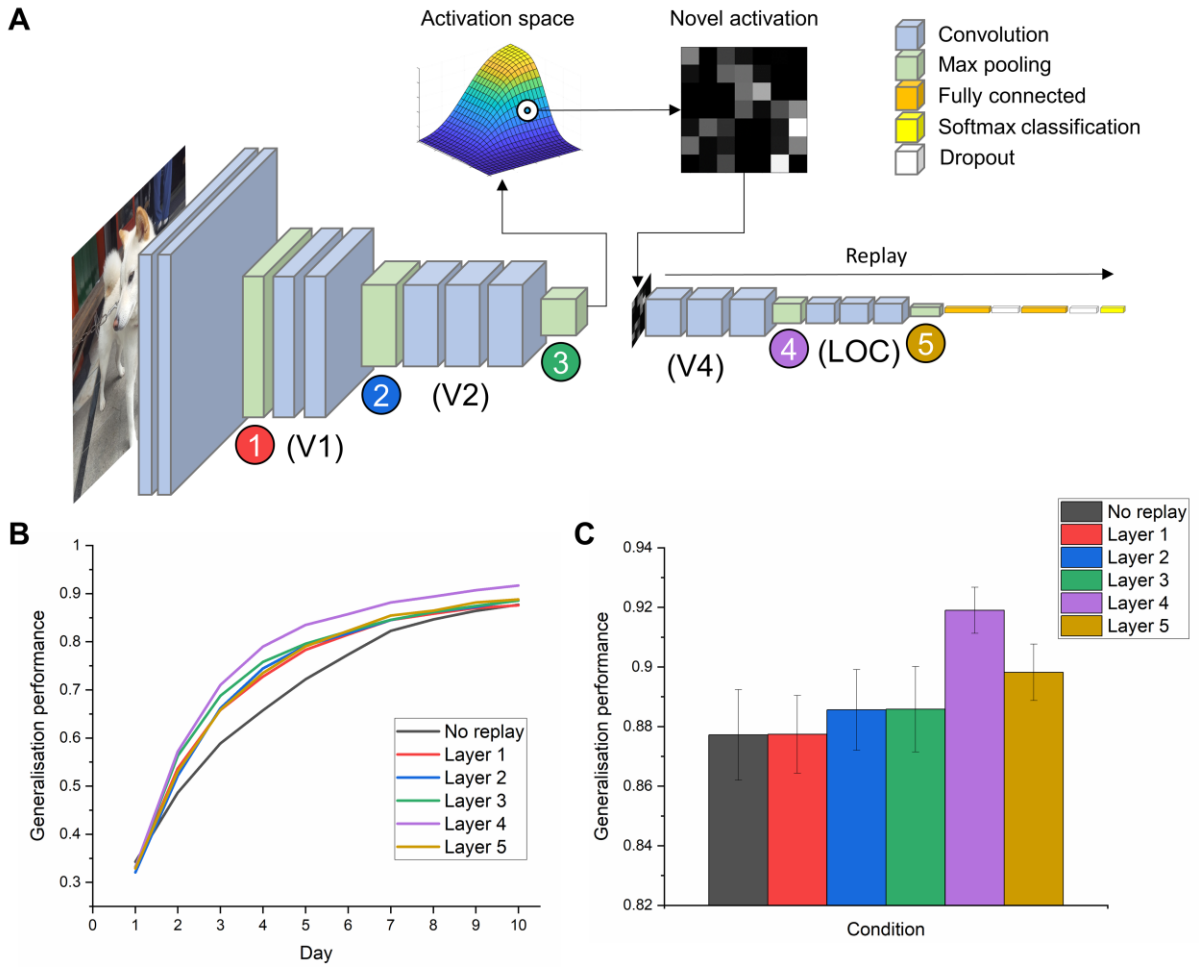
3.1 Localising where in the ventral visual stream generative replay is likely to enhance generalisation

We first sought to establish where in the visual brain the replay of category knowledge might be most effective in helping to generalise to new experiences, as the functional relevance of replay observed in many different brain regions has yet to be established. To simulate the replay process, we used a DCNN called VGG-16, which is already experienced at recognising real-world objects as it has learned to categorise 1000 categories from over one million naturalistic photographs (Simonyan & Zisserman, 2014). Like humans, it can generalise to new situations, and correctly identify the category of an exemplar it has never seen before. It can therefore be viewed as approximating key aspects of a mature visual brain that can support the learning of new categories. Humans readily learn new categories all the time, using previous visual representations to extract useful features such as colour, texture and shape across multiple experiences with an object. VGG-16 emulates this process by using the equivalent building blocks of its own visual experience to extract the key features of objects contained in photographs. Therefore, to simulate new category learning in humans, we tasked this network with learning 10 new categories of objects it has never encountered before. To obtain a baseline measure of how the network would perform without replay, the network learned these 10 new categories in the absence of offline replay. This can be thought of as a human learning new categories in a lab experiment over the course of a single day, without any opportunity to sleep and

consolidate this information in between training blocks. Next, we implemented memory replay. We considered it unrealistic that the human brain could store and replay every single category exemplar it has experienced. Alternatively, humans readily abstract, and are quick to recognise a prototype, or "typical" concept which is representative of category members they have seen (Posner & Keele, 1968), and this process is facilitated by an increased number of experiences (Donald, Joseph, Don, David, & Steven, 1973). Ultimately, this process is important because having a mental prototype helps us to differentiate between categories (Reed, 1972). We therefore deemed it more feasible, efficient, and realistic that humans replay prototypical representations of a category which have been abstracted across learning. We assume, based on neuroimaging studies, that the category prototypes are inherited from higher level regions such as the hippocampus and prefrontal cortex (Bowman et al., 2020), regions which facilitate the learning of concepts (Mack, Love, & Preston, 2018) and imagination (Hassabis, Kumaran, & Maguire, 2007; Mack, Preston, & Love, 2020) of concepts. For the purposes of these experiments, we mimic the function of these higher brain regions in generating prototypical concepts by capturing the "typical" activation of the network for that category and sampling from this gist-like representation to create novel, abstracted representations for replay (Fig 1A). Most replay representations were lower resolution than those during learning (see Materials and Methods) for computational efficiency and to reflect the notional nature of mental imagery.

We simulated generative replay from different layers in the DCNN, equivalent to different brain regions along the ventral stream. Specifically, we trained the network over 10 epochs, corresponding to 10 days of learning, 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 generalise to new exemplars of the categories over the course of learning, and Fig 1C shows the final best performing models in each replay condition. There is a differential benefit of replay throughout the network, where replay in the early layers yields is of limited benefit, whereas replay in the later layers boosts generalisation performance to a greater degree. This suggests that early visual areas in the brain do not store sufficiently complex category-specific representations,

359 curtailing the effectiveness of generated replay representations, whereas higher-level regions such as
360 the lateral occipital cortex are better positioned to support the generation of novel, prototypical
361 concepts which accelerates learning in the absence of real experience and helps us to generalise to
362 new situations.
363



364

365 **Fig 1. The effects of generative replay from different layers of a model of the human ventral visual**
366 **stream on generalisation to new exemplars.** (A) The VGG-16 network simulates the brain's visual
367 system by looking at photographs and extracting relevant features to help categorise the objects
368 within. We trained this network on 10 new categories of objects it had not seen before. In between
369 learning episodes, akin to sleep-facilitated consolidation in humans, we implemented offline memory
370 replay as a generative process. In other words, the network "imagined" new examples of a category
371 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 mid-point in the network, but all five locations where replay was implemented are indicated by the coloured circles. The brain regions corresponding approximately to each network stage, 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 generalise 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. (D) 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 generalisation performance is an advanced layer which bears a functional correspondence to the lateral occipital cortex (LOC) in humans. The benefits of replay from other locations were less pronounced, with the earliest layer showing the least benefit to generalisation.

3.2 Tracking the benefits of replay across learning

Humans encounter new environments throughout their lives, and novel categories which they wish to learn. This knowledge is accumulated and refined across multiple experiences, forming a learning curve for each category. Experiments have focused on the replay of very recently learned information, therefore it is not clear at what point in this learning curve replay is most effective. One could consider replay of recently learned information to be more adaptive, for example, one might want to rapidly consolidate the memory of a plant from which one ate a poisonous berry as one does not want to repeat that experience. Alternatively, generative replay may be less effective for newly encountered categories because there are insufficient experiences from which to adequately extract the underlying prototype. This is a challenging question to address in human experiments, but simulation in an artificial neural network provides an alternate avenue of investigation. In the second experiment, we

extended training to 30 days of experience, interleaved with nights of offline generative replay to simulate learning over longer timescales (Fig 2A). Guided by the results of experiment one, we implemented replay from an advanced layer corresponding to the lateral occipital cortex. In Figure 2D, we show that offline generative replay is most effective at improving generalisation to new exemplars at the earliest stages of learning. This suggests replay facilitates rapid generalisation, which maximises performance given a limited set of experiences with a category.

While establishing that generative replay, or imagining new instances of a category during offline periods, was highly effective in helping to generalise to new category exemplars, 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. As can be seen in Fig 2B, generative replay was comparably effective to veridical replay of experience in consolidating memory, despite being entirely imagined from the networks prior experience. This is despite being a low-resolution gist-like representation, perhaps akin to dreaming about unusual blends of experiences during sleep. This provides compelling support for the hypothesis that generative replay is the most likely form of category replay in humans, as it is vastly more efficient to imagine new concepts from an extracted prototype.

While the aforementioned results show the benefits of replay under optimal conditions where humans encounter the same categories every day, 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 one day, and substituted all subsequent learning experiences with offline replay, termed "continuous replay" (Fig 2A). This is equivalent to a human learning a new category in a one-time lab experiment, and replaying this experience during rest and sleep for the following month. Despite the absence of further exposure to the actual objects, we found the network could increase its generalisation accuracy from 32% to

83% purely by replaying imagined instances of concepts it has partially learned. This may partly account for human's ability to quickly learn from limited experience. However, it also reveals that replayed representations are 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 can be thought of as a constantly evolving "snapshot" of what has been learned so far about that category.

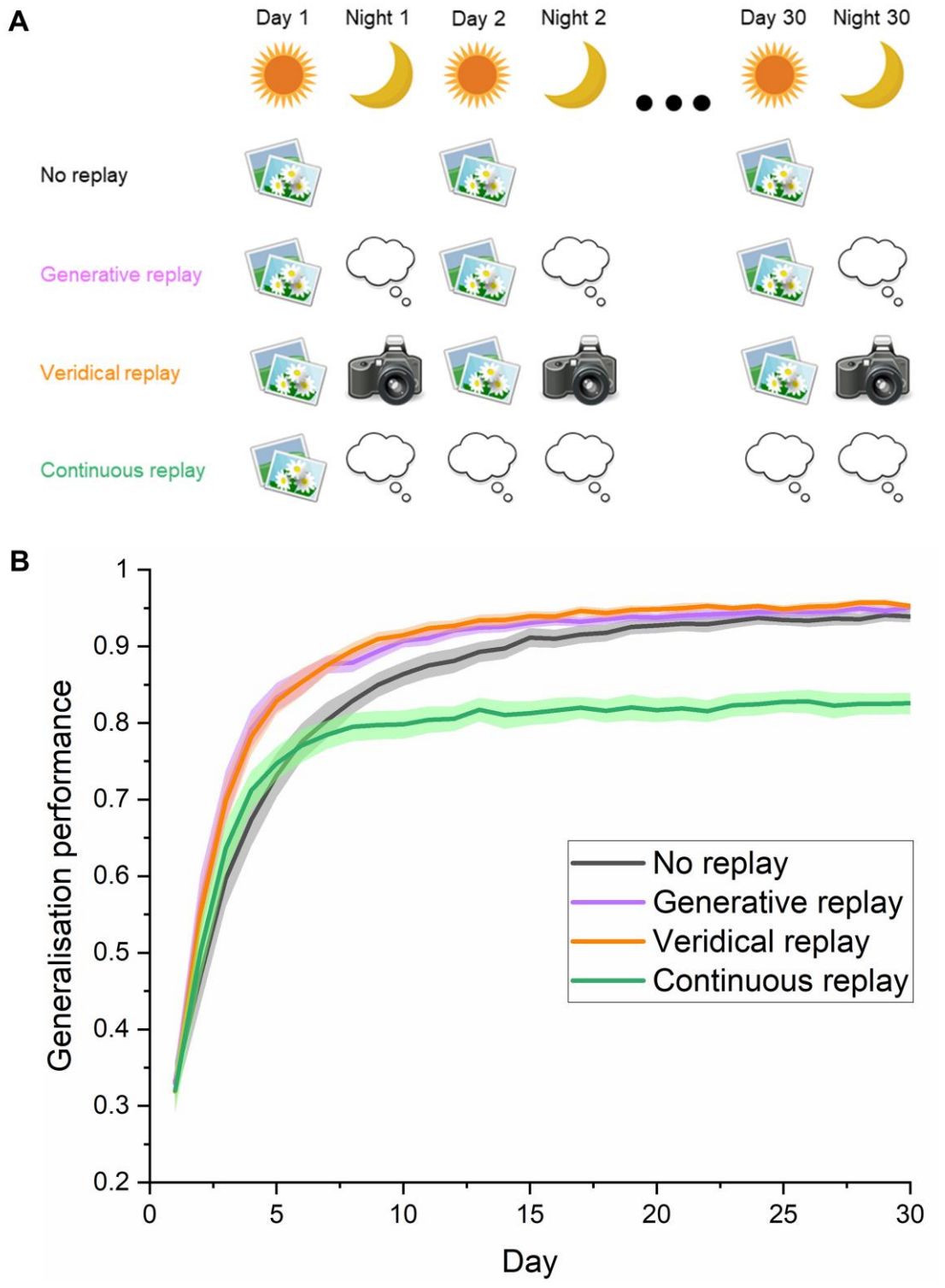


Fig 2. The facilitatory effects of memory replay across category learning. We simulate the long-term consolidation of category memory by extending training to 30 days. (A) Schematic showing the different experimental conditions. "No replay" involves the model of the visual system learning the 10 new 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 equivalent to the lateral occipital cortex. "Veridical replay" tests 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 maximum benefit afforded by replay given only brief exposure to a category. (B) The ability of the network to generalise to new exemplars of a category during each day throughout the learning process. Generalisation performance is measured by the proportion (+/- S.E.M) of correctly recognised test images across 10 models. Generative replay maximally increases performance early in training, suggesting it is critical for new learning and recent memory consolidation. Despite being comprised of internally generated fictive experiences, generative replay was comparably effective to veridical replay throughout the learning process, rendering it an attractive, efficient and more realistic solution to memory consolidation which does not involve remembering all experiences. Continuous replay after just one day of learning substantially improved generalisation performance, but never reached the accuracy levels of networks which engaged in further learning. Replay can therefore compensate for sparse experience to a significant degree, however its limitations also reveal generative replay to be dynamic process, whereby replay representations are informed and improved in tandem with ongoing interleaved learning.

3.3 Determining how the brain might select experiences for replay

Memory consolidation favours weakly-learned information, with a tendency to replay fragile memories more often (Schapiro et al., 2018). How the brain targets these vulnerable representations remains a mystery. Memory replay throughout the brain is triggered by hippocampal activity (Zhang et al., 2018), and given the role of the hippocampus in the generation of prototypes (Bowman et al., 2020), it is likely the hippocampus selects categories for generative 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

463 categories for replay in equal number, however to simulate the autonomous nature of replay selection
464 in the brain, we supplemented our model of the ventral visual stream with a small reinforcement
465 learning network, assuming the theoretical role of the hippocampus in deciding what to replay (Fig
466 3A). The hippocampal model could choose one of the 10 categories to replay, and received a reward
467 from the main network for that action, based on the improvement in network performance.
468 Categories associated with a high reward were more likely to be subsequently replayed, therefore the
469 hippocampal side network could learn through trial and error which categories to replay more often
470 in the cortical network.

471 We trained our model of the visual system on 10 novel categories, implementing replay during
472 offline periods as before, and compared its generalisation performance with that of the dual
473 interactive hippocampal-cortical model. In terms of overall accuracy, both approaches performed
474 similarly throughout training (Fig 3B). However, the reinforcement learning network which simulated
475 the hippocampal replay systematically selected categories which were originally relatively weakly
476 learned more often (Fig 3C), which resulted in their selective improvement. However, this came at a
477 cost, with originally well-learned categories being replayed less often and a drop in their generalisation
478 accuracy. We propose therefore that such a reinforcement learning process may underlie the
479 "rebalancing" of experience in the brain, and that replay helps to compensate for the fact that some
480 categories are more difficult to learn than others.

481

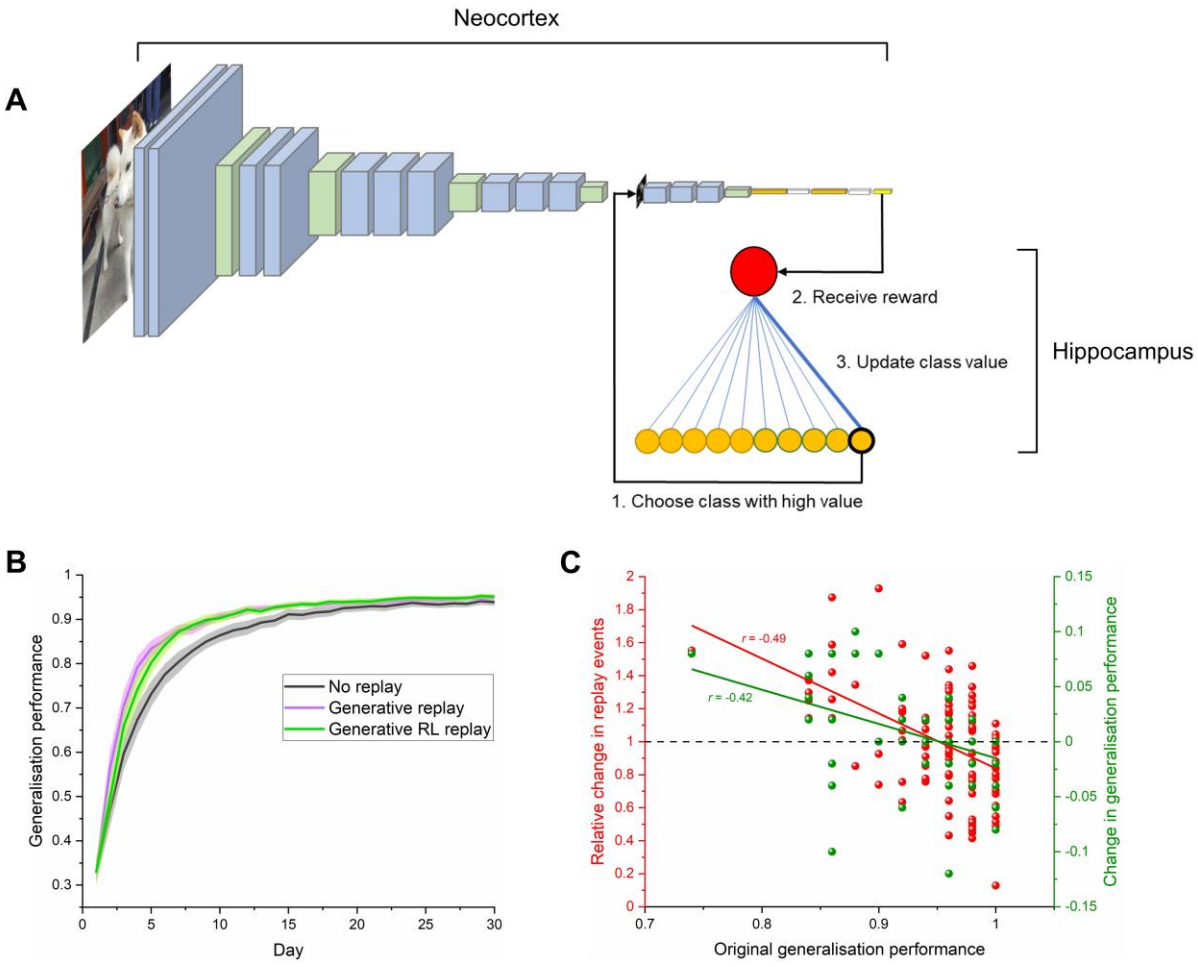


Fig 3. Replay as a reinforcement learning process simulates the brain's tendency to consolidate weaker knowledge. (A) Replay in a model of the visual system is controlled by a reinforcement learning (RL) network akin to the hippocampus. The RL network selects one of 10 categories to replay through the visual system and receives a reward based on the improved performance, learning through trial and error which categories to replay. (B) Overall generalisation performance on new category exemplars was similar for both generative replay and generative replay controlled by a reinforcement learning network. Generalisation performance represents mean accuracy (+/- S.E.M) on test images across 10 models which each learned 10 new categories. (C) The RL network learns to replay categories which were originally more difficult for the visual system, and improves their accuracy. This effectively "rebalances" memory such that category knowledge is more evenly distributed, and offers a candidate mechanism as to how the brain chooses weakly learned

494 information for replay. Plotted values represent the 100 categories across 10 models. A proportion
495 of the generalisation performance values are overlapping.
496

497 4. Discussion

498 We simulated the consolidation of category knowledge in a large-scale neural network model which
499 closely mirrors the form and function of the human ventral visual system, by replaying prototypical
500 representations thought to be formed and initiated by the hippocampus. The notion that replay of
501 visual experiences might be generative in nature has been suggested by limited-capacity models which
502 have been trained on low-resolution photographic images (van de Ven et al., 2020). However, our
503 results using a realistic model of the visual brain represent the most compelling evidence to date that
504 humans are unlikely to replay experiences verbatim during rest and sleep to improve category
505 knowledge, and are more likely to replay novel, imagined instances instead. In addition, the large
506 number (117,000) of high-resolution complex naturalistic images we used for training in this
507 experiment reflected real-world learning and facilitated the extraction of gist-like features. While
508 empirical evidence exists that humans replay novel sequences of stimuli (Liu et al., 2019), our work
509 suggests that the brain goes further and uses learned features of objects to construct entirely fictive
510 experiences to replay. We speculate that this creative process is particularly important for the
511 consolidation of category knowledge as opposed to the replay of episodic memory (Deuker et al.,
512 2013; Schapiro et al., 2018; Zhang et al., 2018), because of the requirement to abstract prototypical
513 features and use these to generalise to new examples of a category. We propose that generative
514 replay confers additional advantages such as constituting less of a burden on memory resources, as
515 not all experiences need to be remembered. Further, our replay representations were highly effective
516 in consolidating category knowledge despite being down-sampled, and these compressed, low-
517 resolution samples would reduce storage requirements further. Perhaps the most convincing
518 demonstration in our simulations that category replay in the brain likely adopts this compressed,
519 prototypical format is that it aided generalisation to a similar degree as the exact veridical replay of

520 experience in boosting generalisation performance. Our findings therefore prompt a
521 reconceptualization of the nature of consolidation-related replay in humans, that it is not only
522 generative, but also low resolution or "blurry", as is the case with internally generated imagery in
523 humans (Giusberti, Cornoldi, De Beni, & Massironi, 1992; Lee, Kravitz, & Baker, 2012). In fact, the kind
524 of replay we propose here may be the driving force behind the transformation of memory into a more
525 schematic, generalised form which preserves regularities across experiences while allowing unique
526 elements of experience to fade (Love & Medin, 1998; Sweegers & Talamini, 2014; Winocur &
527 Moscovitch, 2011). The challenge for future empirical studies in humans to confirm our hypothesis,
528 will be to decode prototypical replay representations during rest and sleep.

529 Simulating replay in a human-like network also allowed us to answer a question not currently
530 tractable in neuroimaging studies: where in the visual stream is replay functionally relevant to
531 consolidation? In a prior simulation of replay in a neural network, van de Ven et al. (2020)
532 demonstrated generative replay could attenuate forgetting when performed after the final
533 convolutional layer, but its effectiveness was not compared to earlier layers, and the network
534 employed, consisting of five convolutional layers, did not reflect the structure of the human visual
535 system. Deeper networks, such as the one used here, consisting of 23 layers in total, organised into
536 five blocks of convolutional layers, not only extract useful category features from naturalistic images,
537 but representations in network layers can be mapped directly on to specific brain regions along the
538 ventral visual stream (Devereux et al., 2018; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte,
539 2014). In keeping with our observation that low-resolution, coarse, schematic replay was effective in
540 helping the network to generalise, we found the most effective location for replay to be in the most
541 advanced layers of the network, layers which are less granular in their representations. This
542 approximately corresponds to the lateral occipital cortex in humans, a region which represents more
543 complex, high-level features (Güçlü & van Gerven, 2015). In contrast, generative replay from the
544 earliest layers corresponding to early visual cortex was less effective. These layers are sensitive to low-
545 level visual features such as contrast, edges and colour, therefore generating samples from these

546 layers will yield rudimentary-level category-specific information, which are of limited utility for replay
547 and generalisation. High-level representations on the other hand, may contain more unique
548 combinations and abstractions of these lower-level features. We also found replay from the
549 penultimate layer was more effective than the final layer, suggesting the optimal replay location
550 represents a balance between the presence of sufficiently complex category information and the
551 number of downstream neuronal weights available to be updated based on replaying these features.
552 These findings prompt a re-evaluation of the functional relevance of replay in early visual cortices in
553 both animals and humans, and generate specific hypotheses for potential perturbation studies to
554 investigate the effects of disruptive stimulation at different stages of the ventral stream during offline
555 consolidation.

556 Our simulations also revealed a phenomenon never before tested in humans, that the
557 effectiveness of replay depends on the stage of learning. We acquire factual information about the
558 world sporadically over time across contexts, for example we may encounter a new species at a zoo
559 one day, and subsequently see the same animal on a wildlife documentary, and so on. Ultimately the
560 consolidation of semantic information in the neocortex can take up to years to complete (Manns et
561 al., 2003). However, our simulations show that replay is most beneficial during the initial encounters
562 with a novel category, when we are still working out its identifiable features and have not yet learned
563 to generalise perfectly to unseen instances. It is therefore likely humans replay a category less and
564 less with increasing familiarity, and there is some support for this idea in the animal literature (Giri et
565 al., 2019). We speculate that the enhanced effectiveness for recent memories may have an adaptive
566 function, allowing us to generalise quickly with limited information. In fact, our simulations showed
567 that after a single learning episode, replay can compensate substantially for an absence of subsequent
568 experience. Our results provide novel hypotheses for human experiments, testing for an interaction
569 between the stage of category learning and the extent of replay. The fact that replay early in the
570 learning process was more effective provides further support for our proposal that vague, imprecise
571 replay events are useful for generalisation, 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 non-converged 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 (Drosopoulos et al., 2007; Kuriyama et al., 2004; 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, Eban, & Frank, 2017), although the content of this neural dialogue is not known. Our simulations suggest that the hippocampus could 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, where generalisation performance was improved for weakly learned items, and attenuated for items which were strongly learned. This reorganisation of knowledge 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, Nitz, McNaughton, & Barnes, 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 and empirical demonstrations of how items get selected for replay within a reinforcement learning framework, such as the "tagging" of items that elicit a large prediction error during the learning phase (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 (Liu et al., 2021; Mattar & Daw, 2018). 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 (Drosopoulos et al., 2007;

598 Kuriyama et al., 2004; McDevitt et al., 2015; Schapiro et al., 2018). Our interactive networks suggest
599 that offline reinforcement learning could account for the selection of weakly-learned knowledge
600 during the replay process itself, and future experiments could assess whether our models choose the
601 same categories for replay as humans when trained on the same stimuli.

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

610

611 **Acknowledgements:** We thank the Love Lab for helpful discussions on this project. This work was
612 supported by the National Institutes of Health (grant number 1P01HD080679); the Royal Society
613 (grant number 183029); and the Wellcome Trust (grant number WT106931MA) to B.C.L. The funders
614 had no role in study design, data collection and analysis, decision to publish, or preparation of the
615 manuscript.

616

617 **Conflicts of Interests:** The authors have declared that no conflicts of interests exist.

618

619 **Author Contributions:** D.N.B: Conceptualization, methodology, software, data curation,
620 investigation, formal analysis, visualization, writing-original draft preparation, writing-review &
621 editing. B.C.L.: Conceptualization, methodology, resources, funding acquisition, supervision, writing-
622 review & editing.

623

Data and Code Availability: The code, environment, and additional information required to run the simulations is available at https://github.com/danielbarry1/replay.git and in the supplementary information. All relevant data in the paper is available at https://doi.org/10.6084/m9.figshare.14208470.

5. References

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *Elife, 9*. doi:10.7554/eLife.59360

de Voogd, L. D., Fernández, G., & Hermans, E. J. (2016). Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. *Neuroimage, 134*, 563-572. doi:10.1016/j.neuroimage.2016.04.026

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248-255.

Deuker, L., Olligs, J., Fell, J., Kranz, T. A., Mormann, F., Montag, C., . . . Axmacher, N. (2013). Memory consolidation by replay of stimulus-specific neural activity. *The Journal of Neuroscience, 33*(49), 19373-19383. doi:10.1523/jneurosci.0414-13.2013

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci Rep, 8*(1), 10636. doi:10.1038/s41598-018-28865-1

Donald, H., Joseph, C., Don, C., David, G., & Steven, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101*(1), 116-122. doi:10.1037/h0035772

Drosopoulos, S., Windau, E., Wagner, U., & Born, J. (2007). Sleep enforces the temporal order in memory. *PLoS One, 2*(4), e376. doi:10.1371/journal.pone.0000376

Dudai, Y. (2012). The restless engram: consolidations never end. *Annu Rev Neurosci, 35*, 227-247. doi:10.1146/annurev-neuro-062111-150500

Eichenlaub, J.-B., Jarosiewicz, B., Saab, J., Franco, B., Kelemen, J., Halgren, E., . . . Cash, S. S. (2020). Replay of learned neural firing sequences during rest in human motor cortex. *Cell Rep, 31*(5), 107581. doi:10.1016/j.celrep.2020.107581

Ekman, M., Kok, P., & de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nat Commun, 8*(1), 15276. doi:10.1038/ncomms15276

Fan, Y., Tian, F., Qin, T., Li, X.-Y., & Liu, T.-Y. (2018). Learning to teach. *arXiv preprint arXiv:1805.03643*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends Cogn Sci, 3*(4), 128-135. doi:10.1016/S1364-6613(99)01294-2

Friedrich, M., Wilhelm, I., Born, J., & Friederici, A. D. (2015). Generalization of word meanings during infant sleep. *Nat Commun, 6*(1), 6004. doi:10.1038/ncomms7004

Girardeau, G., Inema, I., & Buzsáki, G. (2017). Reactivations of emotional memory in the hippocampus–amygdala system during sleep. *Nat Neurosci, 20*(11), 1634. doi:10.1038/nn.4637

Giri, B., Miyawaki, H., Mizuseki, K., Cheng, S., & Diba, K. (2019). Hippocampal reactivation extends for several hours following novel experience. *The Journal of Neuroscience, 39*(5), 866-875. doi:10.1523/JNEUROSCI.1950-18.2018

667 Giusberti, F., Cornoldi, C., De Beni, R., & Massironi, M. (1992). Differences in vividness ratings of
668 perceived and imagined patterns. *British Journal of Psychology, 83*(4), 533-547.
669 doi:10.1111/j.2044-8295.1992.tb02457.x
670 Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of
671 neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005-
672 10014. doi:10.1523/JNEUROSCI.5023-14.2015
673 Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not
674 a simple function of experience. *Neuron, 65*(5), 695-705. doi:10.1016/j.neuron.2010.01.034
675 Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis
676 of episodic memory. *The Journal of Neuroscience, 27*(52), 14365-14374.
677 doi:10.1523/JNEUROSCI.4549-07.2007
678 Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia
679 cannot imagine new experiences. *Proc Natl Acad Sci U S A, 104*(5), 1726-1731.
680 doi:10.1073/pnas.0610561104
681 Hayes, T. L., Krishnan, G. P., Bazhenov, M., Siegelmann, H. T., Sejnowski, T. J., & Kanan, C. (2021).
682 Replay in Deep Learning: Current Approaches and Missing Biological Elements.
683 *https://arxiv.org/abs/2104.04132*.
684 He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving deep into rectifiers: Surpassing human-level*
685 *performance on imagenet classification.* Paper presented at the Proceedings of the IEEE
686 international conference on computer vision.
687 Hermans, E. J., Kanen, J. W., Tambini, A., Fernández, G., Davachi, L., & Phelps, E. A. (2017).
688 Persistence of amygdala–hippocampal connectivity and multi-voxel correlation structures
689 during awake rest after fear learning predicts long-term expression of fear. *Cereb Cortex,*
690 *27*(5), 3028-3041. doi:10.1093/cercor/bhw145
691 Horváth, K., Liu, S., & Plunkett, K. (2016). A daytime nap facilitates generalization of word meanings
692 in young toddlers. *Sleep, 39*(1), 203-207. doi:10.5665/sleep.5348
693 Jafarpour, A., Fuentemilla, L., Horner, A. J., Penny, W., & Duzel, E. (2014). Replay of very early
694 encoding representations during recollection. *J Neurosci, 34*(1), 242-248.
695 doi:10.1523/JNEUROSCI.1865-13.2014
696 Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus
697 during sleep. *Nat Neurosci, 10*(1), 100-107. doi:10.1038/nn1825
698 Kemker, R., & Kanan, C. (2017). Fearnet: Brain-inspired model for incremental learning. *arXiv*
699 *preprint arXiv:1711.10563*.
700 Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may
701 explain IT cortical representation. *PLoS Comput Biol, 10*(11), e1003915.
702 doi:10.1371/journal.pcbi.1003915
703 Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*
704 *arXiv:1412.6980*.
705 Kuriyama, K., Stickgold, R., & Walker, M. P. (2004). Sleep-dependent learning and motor-skill
706 complexity. *Learn Mem, 11*(6), 705-713. doi:10.1101/lm.76304
707 Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-
708 world objects. *Neuroimage, 59*(4), 4064-4073. doi:10.1016/j.neuroimage.2011.10.055
709 Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously
710 reorganizes experience. *Cell, 178*(3), 640-652.e614. doi:10.1016/j.cell.2019.06.012
711 Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is
712 associated with efficient nonlocal learning. *Science, 372*(6544). doi:10.1126/science.abf1357
713 Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Aaai/iaai*, 671-676.
714 Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The
715 hippocampus and concept formation. *Neurosci Lett, 680*, 31-38.
716 doi:10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nat Commun, 11*(1), 46. doi:10.1038/s41467-019-13930-8

Manns, J. R., Hopkins, R. O., & Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron, 38*(1), 127-133. doi:10.1016/S0896-6273(03)00146-6

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *21*(11), 1609-1617. doi:10.1038/s41593-018-0232-z

McDevitt, E. A., Duggan, K. A., & Mednick, S. C. (2015). REM sleep rescues learning from interference. *Neurobiol Learn Mem, 122*, 51-62. doi:10.1016/j.nlm.2014.11.015

Michelmann, S., Staresina, B. P., Bowman, H., & Hanslmayr, S. (2019). Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nature Human Behaviour, 3*(2), 143-154. doi:10.1038/s41562-018-0491-4

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife, 7*, e32548.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., & Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat Neurosci, 12*(7), 919-926. doi:10.1038/nn.2337

Poe, G. R., Nitz, D. A., McNaughton, B. L., & Barnes, C. A. (2000). Experience-dependent phase-reversal of hippocampal neuron firing during REM sleep. *Brain Res, 855*(1), 176-180. doi:10.1016/S0006-8993(99)02310-0

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353. doi:10.1037/h0025953

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*(3), 382-407. doi:10.1016/0010-0285(72)90014-X

Remy, P. (2020). Keract: A library for visualizing activations and gradients. *GitHub repository*. Retrieved from https://github.com/philipperemy/keract

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., . . . Kording, K. P. (2019). A deep learning framework for neuroscience. *Nat Neurosci, 22*(11), 1761-1770. doi:10.1038/s41593-019-0520-2

Robins, A. (1995). Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connection Science, 7*(2), 123-146. doi:10.1080/09540099550039318

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211-252. doi:10.1007/s11263-015-0816-y

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Sci Rep, 7*(1), 14869. doi:10.1038/s41598-017-12884-5

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat Commun, 9*(1), 3920. doi:10.1038/s41467-018-06213-1

Schönauer, M., Alizadeh, S., Jamalabadi, H., Abraham, A., Pawlizki, A., & Gais, S. (2017). Decoding material-specific memory reprocessing during sleep in humans. *Nat Commun, 8*, 15404. doi:10.1038/ncomms15404

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-Like? *bioRxiv*, 407007. doi:10.1101/407007

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., . . . Behrens, T. (2021). Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *bioRxiv*, 2021.2006.2006.447249. doi:10.1101/2021.06.06.447249

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15*(1), 1929-1958.

Staresina, B. P., Alink, A., Kriegeskorte, N., & Henson, R. N. (2013). Awake reactivation predicts memory in humans. *Proc Natl Acad Sci U S A, 110*(52), 21159-21164. doi:10.1073/pnas.1311989110

Sweegers, C. C. G., & Talamini, L. M. (2014). Generalization from episodic memories across time: A route for semantic knowledge acquisition. *Cortex, 59*, 49-61. doi:10.1016/j.cortex.2014.07.006

Tambini, A., & Davachi, L. (2013). Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc Natl Acad Sci U S A, 110*(48), 19591-19596. doi:10.1073/pnas.1308499110

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron, 65*(2), 280-290. doi:10.1016/j.neuron.2010.01.001

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

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265*(5172), 676-679. doi:10.1126/science.8036517

Wimmer, G. E., Liu, Y., Vehar, N., Behrens, T. E. J., & Dolan, R. J. (2020). Episodic memory retrieval success is associated with rapid replay of episode content. *Nat Neurosci, 23*(8), 1025-1033. doi:10.1038/s41593-020-0649-z

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *J Int Neuropsychol Soc, 17*(5), 766-780. doi:10.1017/S1355617711000683

Wittkuhn, L., & Schuck, N. W. (2021). Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nat Commun, 12*(1), 1795. doi:10.1038/s41467-021-21970-2

Zhang, H., Fell, J., & Axmacher, N. (2018). Electrophysiological mechanisms of human memory consolidation. *Nat Commun, 9*(1), 4103. doi:10.1038/s41467-018-06553-y

808 Supplementary table 1: List of ImageNet classes by model

Model 1	n12360108 begonia
	n02822579 bedstead bedframe
	n02427724 waterbuck
	n03098688 control room
	n02944075 camisole
	n01603600 waxwing
	n03196598 digital display alphanumeric display
	n02848216 blade
	n07712856 tortilla chip
	n03592669 jalousie
Model 2	n11853356 Christmas cactus Schlumbergera buckleyi Schlumbergera baridgesii
	n04177820 settle settee
	n03904183 pedestrian crossing zebra crossing
	n04355511 sundress
	n03487444 hand lotion
	n12899752 angel's trumpet Brugmansia suaveolens Datura suaveolens
	n12655869 raspberry raspberry bush
	n12948053 common European dogwood red dogwood blood-twig pedwood Cornus sanguinea
	n02869737 bongo bongo drum
	n02415253 Dall sheep Dall's sheep white sheep Ovis montana dalli
Model 3	n03375575 foil
	n03082807 compressor
	n03262932 easy chair lounge chair overstuffed chair
	n02047614 puffin
	n03317788 faience
	n09475044 wasp's nest wasps' nest hornet's nest hornets' nest
	n11784497 jack-in-the-pulpit Indian turnip wake-robin Arisaema triphyllum Arisaema atrorubens
	n03941231 pinata
	n02813399 bay window bow window
	n04544325 wainscoting wainscotting
Model 4	n03993053 potty seat potty chair
	n04082886 reticle reticule graticule
	n03421324 garter belt suspender belt
	n03766044 miller milling machine
	n03505504 headscarf
	n12384839 love-in-a-mist running pop wild water lemon Passiflora foetida
	n03619793 kitbag kit bag
	n07600696 candied apple candy apple taffy apple caramel apple toffee apple
	n02068974 dolphin
	n03237992 dressing gown robe-de-chambre lounging robe
Model 5	n02918964 bumper car Dodgem
	n02392824 white rhinoceros Ceratotherium simum Diceros simus

	n01806364 blue peafowl Pavo cristatus
	n02956699 capitol
	n04290079 spun yarn
	n08596076 littoral litoral littoral zone sands
	n02887970 bracelet bangle
	n10635788 sphinx
	n07901457 muscat muscatel muscadel muscadelle
	n07870167 lasagna lasagne
Model 6	n04324387 stockroom stock room
	n04591517 wind turbine
	n02988486 CD-R compact disc recordable CD-WO compact disc write-once
	n04568069 weathervane weather vane vane wind vane
	n04514241 uplift
	n03207835 dishtowel dish towel tea towel
	n13206817 maidenhair maidenhair fern
	n03307792 external drive
	n12666965 cape jasmine cape jessamine Gardenia jasminoides Gardenia augusta
	n12950126 valerian
Model 7	n03986355 portfolio
	n11848479 night-blooming cereus
	n04439712 tinfoil tin foil
	n03160740 damask
	n01612122 sparrow hawk American kestrel kestrel Falco sparverius
	n09206896 arroyo
	n12392549 stinging nettle Urtica dioica
	n02343772 gerbil gerbille
	n07875436 risotto Italian rice
	n02060133 fulmar fulmar petrel Fulmarus glacialis
Model 8	n03655072 legging leging leg covering
	n10738111 unicyclist
	n09270735 dune sand dune
	n03409393 gable gable end gable wall
	n02331046 rat
	n03452267 gramophone acoustic gramophone
	n10105733 forward
	n07911677 cocktail
	n03797182 muffler
	n01563128 warbler
Model 9	n04197110 shipwreck
	n10470779 priest
	n02769290 backhoe
	n03478756 hall
	n04519153 valve
	n04289027 sprinkler
	n02782778 ballpark park

	n03558404 ice skate
	n04138261 satin
	n02700064 alternator
Model 10	n03524150 hockey stick
	n03716966 mandolin
	n02962200 carburetor carburettor
	n03237340 dresser
	n04004210 printed circuit
	n02917377 bullhorn loud hailer loud-hailer
	n07879953 tempura
	n04087826 ribbing
	n02404432 longhorn Texas longhorn
	n07830593 hot sauce

809
810

811
812