A neural network account of memory replay and knowledge consolidation

Short 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 by offline neural reactivation related to past experiences. Category knowledge is learned across multiple experiences and subsequently generalised to new situations. This ability to generalise is promoted by offline consolidation and replay during rest and sleep. However, aspects of replay are difficult to determine from neuroimaging studies alone. Here, we provide a comprehensive account of how category replay may work in the brain by simulating these processes in a neural network which assumed the functional roles of the human ventral visual stream and hippocampus. We showed that generative replay, akin to imagining entirely new instances of a category, facilitated generalisation to new experiences. This invites a reconsideration of the nature of replay more generally, and suggests that replay helps to prepare us for the future as much as remember the past. We simulated generative replay at different network locations finding it was most effective in later layers equivalent to the lateral occipital cortex, and less effective in layers corresponding to early visual cortex, thus drawing a distinction between the observation of replay in the brain and its relevance to consolidation. We modelled long-term memory consolidation in humans and found that category replay is most beneficial for newly acquired knowledge, at a time when generalisation is still poor, a finding which suggests replay helps us adapt to changes in our environment. Finally, we present a novel mechanism for the frequent observation that the brain selectively consolidates weaker information, and showed that a reinforcement learning process in which categories were replayed according to their contribution to network performance explains this well-documented phenomenon, thus reconceptualising replay as an active rather than passive process.

Author Summary

The brain relives past experiences during rest. This process is called "replay" and helps strengthen our memories and promote generalisation. We learn over time to categorise objects, yet how category knowledge is replayed in the brain is not well understood. We used a neural network which behaves like the human visual brain to simulate category replay. We found that allowing the network to "dream" typical examples of a category during "night-time" consolidation was an effective form of replay that helped subsequent recognition of unseen objects, offering a solution for how the human brain consolidates category knowledge. We also found this to be more effective if it took place in advanced layers of the network, suggesting human replay might be most effective in high-level visual brain regions. We also tasked the network with learning to control its own replay, and found it focused on categories that were difficult to learn. This represents the first mechanistic account of why weakly-learned memories in humans show the greatest improvement after rest and sleep. Our approach makes predictions about category replay in the human brain which can inform future experiments, and highlights the value of large-scale neural networks in addressing neuroscientific questions.

1. Introduction

Memory replay refers to the reactivation of experience-dependent neural activity during resting periods. First observed in rodent hippocampal cells during sleep [1], the phenomenon has since been detected in humans during rest [2-6], and sleep [7, 8], 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. However, in this paper we challenge the notion of replay as a passive, memory-preserving process, and propose it is much more dynamic in nature. Using a computational approach, we test hypotheses that 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.

Replay is assumed to constitute the veridical reactivation of past experience. However, there are circumstances in which this may be suboptimal or impractical. For example, a desirable outcome of category replay is to generalise to new experiences rather than recognise past instances, a phenomenon observed after sleep in infants [9, 10]. In addition, although sleep benefits category learning for a limited number of well-controlled experimental stimuli [11], 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 [8] and the generation of prototypical concepts [12], we consider this the most likely form of category replay. The replay of novel [13] and random [14] spatial trajectories have been decoded from hippocampal “place cells” in animals. However, due to the complex nature of category knowledge, detecting such novel replay events from human brain data would be challenging.

The occurrence of replay in humans is associated with subsequent memory [8]. However, establishing a causal relationship between observed neural reactivation and memory consolidation is problematic. Replay has been observed throughout the brain, early in the ventral visual stream [6, 15, 16], in the ventral temporal cortex [17, 18], the medial temporal lobe [5, 19] the amygdala, [3, 20], motor cortex [21] and prefrontal cortex [22]. It is not known if replay in low-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.

Because it can take humans years to learn and consolidate semantic or conceptual knowledge [23], we still do not know how long replay contributes to this process, as neuroimaging studies are limited to a time-span of a day or two. Humans are thought to "reconsolidate" information every time it is retrieved [24], 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 [25], suggesting the brain only replays recently learned, vulnerable information. Answering this question in humans remains a challenge because of the practicalities of tracking replay events for extended periods.

It has been frequently observed that replay and consolidation selectively benefit weakly-learned over well-learned information [5, 26-28], but a candidate mechanism for how this occurs in the brain has not been proposed to date.

Our understanding of replay in the human brain is therefore limited by the difficulty in measuring and perturbing this covert, spontaneous process. However, an alternative approach which can address these outstanding questions, is to harness the recent considerable advances in artificial neural networks. While replay has been previously simulated in smaller-scale networks [29-31], in order to make direct comparisons with the human brain, we simulated learning and replay in a deep convolutional neural network (DCNN) which mirrors the brain's layered structure and representations [32, 33] and approaches human-level recognition performance [34]. To simulate new learning in humans, we took a network which has already been trained to successfully categorise 1000 categories of objects in photographs, akin to a fully functional visual system in humans, and tasked it with learning 10 novel categories. 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 replay in the network, akin to human consolidation during sleep. We targeted replay at specific network layers functionally equivalent to different brain regions to make novel predictions about where in the brain replay is causally effective. We evaluated whether prior reports of replay in early visual areas are likely to be relevant to memory consolidation. 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 also simulated "imagined" prototypical replay events and determine whether this was as effective as veridical replay in helping us to generalise to new, unseen experiences, thus supporting our conceptualization of replay as a creative process. We simulated the learning of categories across multiple experiences to make predictions about when in learning replay is likely to be effective in boosting subsequent generalisation performance. We hypothesised that the benefits of replay may be confined to early in the learning curve when novel category knowledge is being acquired. We also tested a mechanism through which the brain selects certain items for replay, adding an auxiliary model (akin to the hippocampus) to the neocortical model, which could autonomously learn the best consolidation strategy, determining what to replay and when. We predicted that this dynamic process would result in the prioritisation of weakly-learned items, in line with behavioural studies of memory consolidation. The overall aim of these experiments was to provide answer questions about memory replay in humans using a model of the human visual ventral stream, and this aim was successfully achieved.

2. Results

2.1 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 [35]. Like humans, it can generalise to new situations, and correctly identify the category of an exemplar it has never seen before. It has achieved a high "Brain-Score" which is a benchmark for how closely a neural network reflects the brain's neural representations and object recognition behaviour in primates [36]. 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 [37], and this process is facilitated by an increased number of experiences [38]. Ultimately, this process is important because having a mental prototype helps us to differentiate between categories [39]. 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 [40], regions which facilitate the learning of concepts [41] and imagination [42, 43] 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 Methods and Models) 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. This suggests that early visual areas in the brain do not contain sufficient category-specific information to form useful replay representations, whereas higher-level regions such as the lateral occipital cortex can support the generation of novel, prototypical concepts which accelerates learning in the absence of real experience and helps us to generalise to new situations.

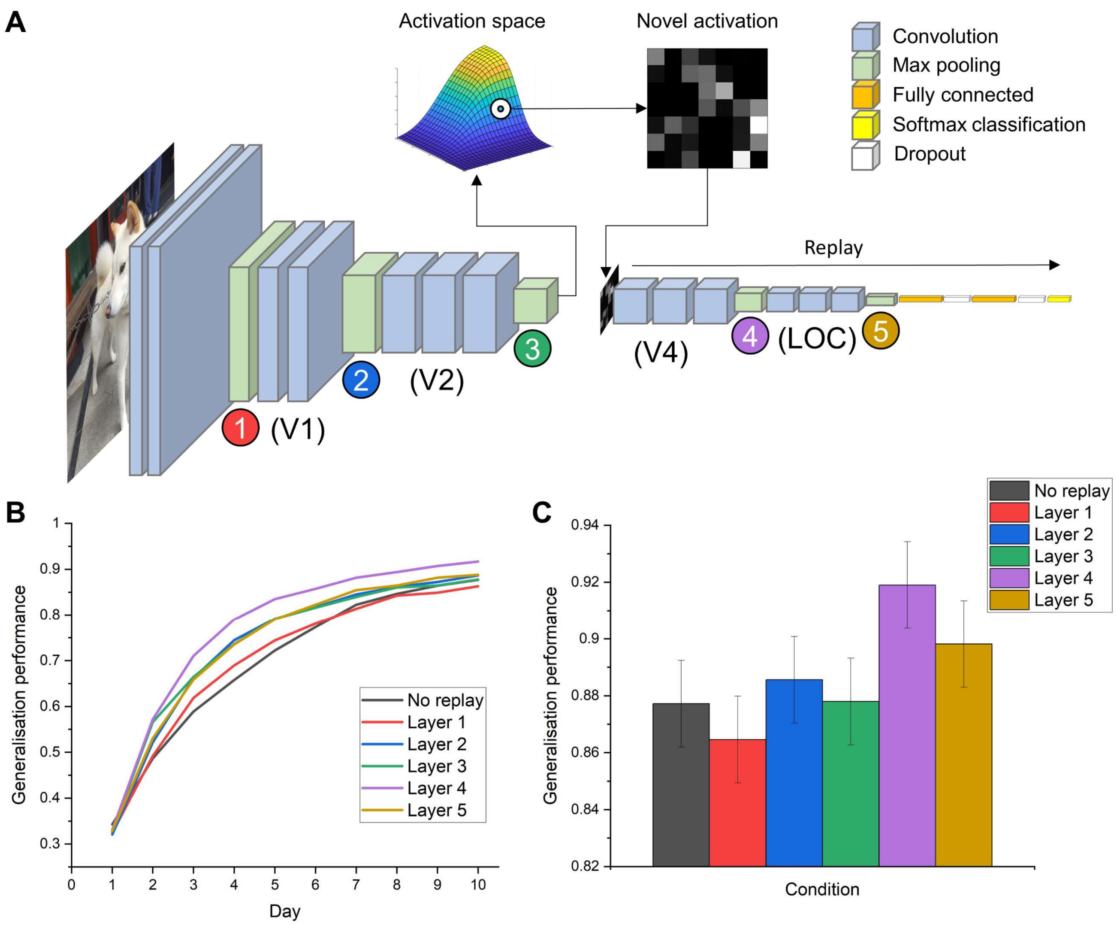

Fig 1. The effects of generative replay from different layers of a model of the human ventral visual stream on generalisation to new exemplars. (A) The VGG-16 network simulates the brain's visual system by looking at photographs and extracting relevant features to help categorise the objects within. We trained this network on 10 new categories of objects it had not seen before. In between learning episodes, akin to 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 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 (32), 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.

2.2 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.

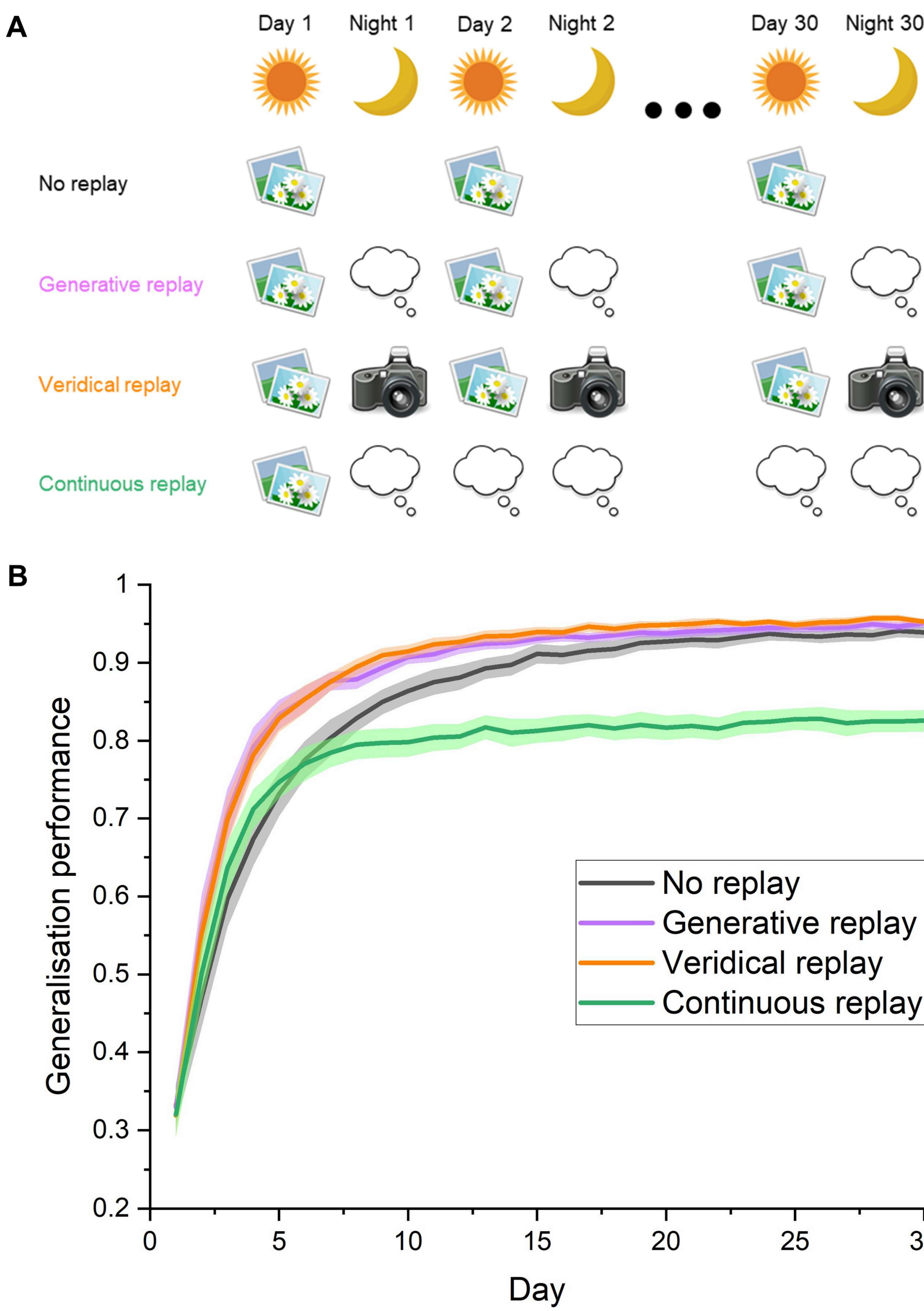

Fig 2. The facilitatory effects of memory replay across category learning. We simulate the long-term consolidation of category memory by extending training to 30 days. (A) Schematic showing the different experimental conditions. "No replay" involves the model of the visual system learning the 10 new 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.

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 this from the same point in the network. As can be seen in Fig 2B, generative replay was as effective as 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.

2.3 Determining how the brain might select experiences for replay

Memory consolidation favours weakly-learned information, with a tendency to replay fragile memories more often [5]. How the brain targets these vulnerable representations remains a mystery. Memory replay throughout the brain is triggered by hippocampal activity [8], and given the role of the hippocampus in the generation of prototypes [40], 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 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 network, assuming the theoretical role of the hippocampus in deciding what to replay (Fig 3A). The hippocampal 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. Categories associated with a high reward were more likely to be subsequently replayed, therefore the hippocampal side network could learn through trial and error which categories to replay more often in the cortical network.

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

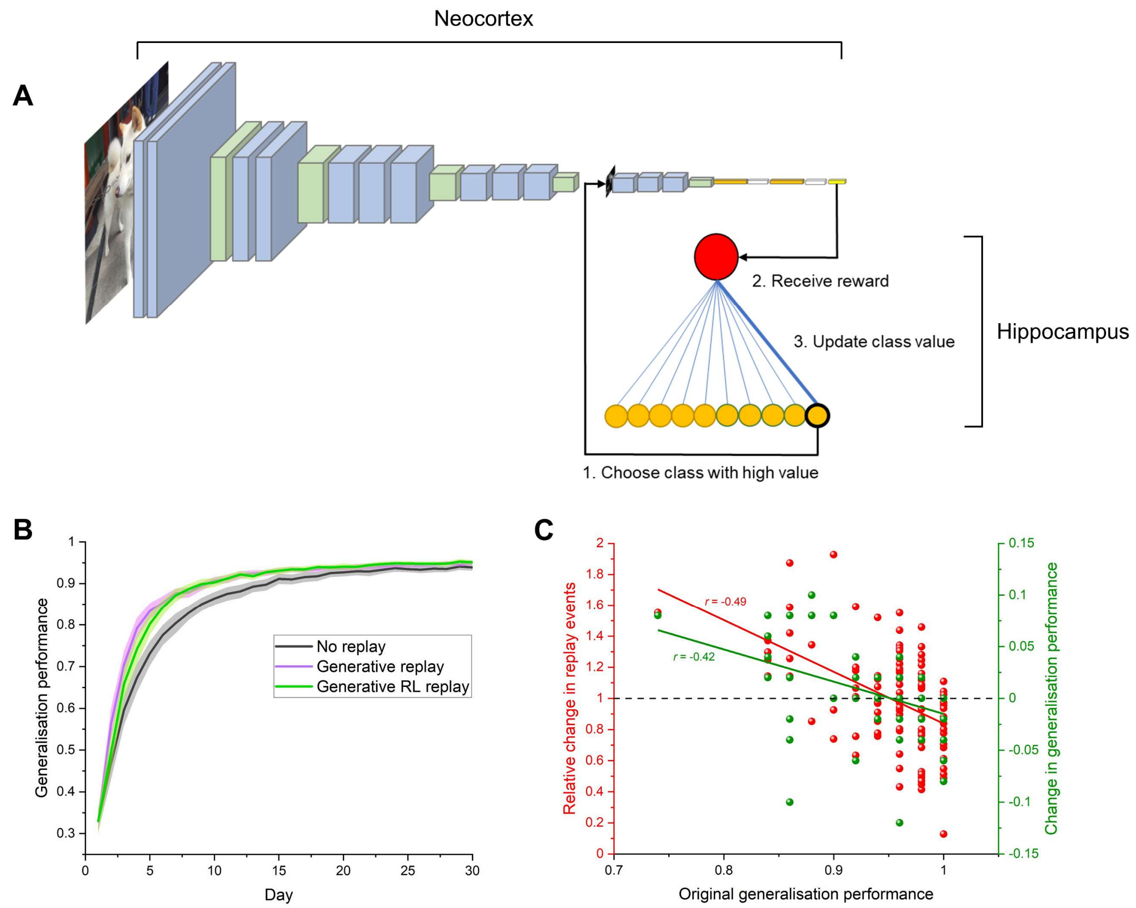

Fig 3. Replay as a reinforcement learning process simulates the brain's tendency to consolidate weaker knowledge. (A) Replay in a model of the visual system is controlled by a reinforcement learning (RL) network akin to the hippocampus. The RL network selects one of 10 categories to replay through the 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 information for replay. Plotted values represent the 100 categories across 10 models. A proportion of the generalisation performance values are overlapping.

3. Discussion

We simulated the consolidation of category knowledge in a large-scale neural network model which closely mirrors the form and function of the human ventral visual system, by replaying prototypical representations thought to be formed and initiated by the hippocampus. The notion that replay might be generative in nature has been suggested by smaller simulations [30, 31], however our results using a realistic model of the visual brain represent the most compelling evidence to date that humans are unlikely to replay experiences verbatim during rest and sleep to improve category knowledge, and are 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 reflected real-world learning and facilitated the extraction of gist-like features. While empirical evidence exists that humans replay novel sequences of stimuli [4], our work suggests that the brain goes further and uses 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 [5, 8, 15], because of the requirement to abstract prototypical features and use these to generalise 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 down-sampled, and these compressed, low-resolution samples would reduce storage requirements further. Perhaps the most convincing demonstration in our simulations that category replay in the brain likely adopts this compressed, prototypical format is that it was as effective as the exact veridical replay of experience in boosting generalisation performance. Our findings therefore prompt a reconceptualization of the nature of 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 [44, 45]. In fact, the kind of replay we propose here may be the driving force behind the transformation of memory into a more schematic, generalised form which preserves regularities across experiences while allowing unique elements of experience to fade [46-48]. The challenge for future empirical studies in humans to confirm our hypothesis, will be to decode prototypical replay representations during rest and sleep.

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 keeping with our observation that low-resolution, coarse, schematic replay was effective in helping the network to generalise, 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 approximately corresponds to the lateral occipital cortex in humans, a region which represents more complex, high-level features [32]. In contrast, generative replay from the earliest layers corresponding to early visual cortex was ineffective, suggesting more precise, fine-grained replay might not be optimal in preparing the brain to recognise novel instances of a category. In addition, these layers are sensitive to low-level visual features such as contrast and edges, which are likely shared across all categories, and therefore do not contain enough distinctive information to be useful for replay or generalisation. High-level representations on the other hand, may contain more unique combinations and abstractions of these lower-level features. This prompts a re-evaluation of the functional relevance of replay in early visual cortices in both animals and humans, and generates 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 one 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 [23]. However, our simulations show that replay is 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 generalise perfectly to unseen instances. It is therefore likely humans replay a category less and less with increasing familiarity, and there is some support for this idea in the animal literature [25]. We speculate that the enhanced effectiveness for recent memories may have an adaptive function, allowing us to generalise 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 generalisation, as the networks imaginary representations at that stage would be an imperfect approximation of the category in question.

Our results also represent the first mechanistic account of how the brain selects weakly-learned information for replay and consolidation [5, 26-28]. The hippocampus triggers replay events in the neocortex [8], with a loop of information back and forth between the two brain areas [49], 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 [50]. 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. Future experiments could assess whether our interactive models choose the same categories for replay as humans when trained on the same stimuli.

In summary, our simulations provide strong evidence that category replay in humans is a generative process which is functionally relevant at advanced stages of the ventral stream. We make testable predictions 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 dialogue 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 [51].

4. Methods and models

4.1 Neural network

To simulate the learning of novel concepts in the brain, and test a number of hypotheses regarding replay, we trained a DCNN on 10 new categories of images. The neural network was VGG-16 [35]. Emulating the extent of real-world learning in humans, this network is trained on a vast dataset of 1.3 million naturalistic photographs known as the ImageNet database [52], which contains recognisable objects from 1000 categories in different contexts much like what humans encounter on a daily basis. The network learns to associate the visual features of an object with its category label, until it can recognise examples of that object which it has never seen before, reflecting the human ability to generalise prior knowledge to new situations. The network takes a photograph's pixels as input, and sequentially transforms this input into more abstract features, similar to the operation of the human ventral visual stream [36]. 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.

This network which has been previously trained on 1000 categories can be thought of as 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 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 [53]. 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.

4.2 Stimuli

Photographic stimuli for new classes were drawn randomly from the larger ImageNet 2011 fall database [54], 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 S1.

4.3 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 locations, or from different angles, and facilitates the extraction of useful features rather than rote memorisation of experience. Augmentation could include up to 20-degree 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 pre-processed in accordance with Simonyan and Zisserman (35). Depending on the experiment, the network was trained for 10 or 30 epochs. We used the Adam optimiser [55] with a learning rate of 0.0003. The training batch size was set to 36. The training objective was to minimise the categorical cross-entropy loss over the 10 classes. Training parameters were optimised based on validation set performance. We report the model's performance metrics from the test set only, which reflects the model's ability to generalise to new stimuli during and after training. Training was performed using TensorFlow version 2.2.

4.4 Replay

Replay was conducted between training epochs, to simulate "days" of learning and "nights" of offline consolidation. We conceptualised replay representations as generative, in other words they represented a prototype of that category never seen before, from a particular point in the network. This represents an alternative to storing every experience in our heads, in that we could replay important knowledge about the world without remembering everything. 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 [56]. For each class separately, a multivariate distribution of activity was created from these activations, representing the unique relationship between units of the layer which were observed for that specific 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 [12, 40].

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 four for blocks one and two, and by two for blocks three and 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 item 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 [30, 31].

4.5 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 [8]. This approach is similar to the "teacher-student" meta-learning framework which has been shown to improve performance in deep neural networks [57]. 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.

Funding information: This research was supported by NIH Grant 1P01HD080679 (https://www.nih.gov/), Royal Society Wolfson Fellowship 183029 (https://royalsociety.org/), and a Wellcome Trust Senior Investigator Award WT106931MA (https://wellcome.org/) held by B.C.L. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

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

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

1. Wilson MA, McNaughton BL. Reactivation of hippocampal ensemble memories during sleep. Science. 1994;265(5172):676-9. doi: 10.1126/science.8036517.
2. Tambini A, Davachi L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(48):19591-6. doi: 10.1073/pnas.1308499110.
3. 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. Cerebral cortex. 2017;27(5):3028-41. doi: 10.1093/cercor/bhw145.
4. Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ. Human replay spontaneously reorganizes experience. Cell. 2019;178(3):640-52.e14. doi: 10.1016/j.cell.2019.06.012.
5. Schapiro AC, McDevitt EA, Rogers TT, Mednick SC, Norman KA. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. Nature communications. 2018;9(1):3920. doi: 10.1038/s41467-018-06213-1.
6. Wittkuhn L, Schuck NW. Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. Nature communications. 2021;12(1):1795. doi: 10.1038/s41467-021-21970-2.
7. Schönauer M, Alizadeh S, Jamalabadi H, Abraham A, Pawlizki A, Gais S. Decoding material-specific memory reprocessing during sleep in humans. Nature communications. 2017;8:15404. doi: 10.1038/ncomms15404.
8. Zhang H, Fell J, Axmacher N. Electrophysiological mechanisms of human memory consolidation. Nature communications. 2018;9(1):4103. doi: 10.1038/s41467-018-06553-y.
9. Friedrich M, Wilhelm I, Born J, Friederici AD. Generalization of word meanings during infant sleep. Nature communications. 2015;6(1):6004. doi: 10.1038/ncomms7004.
10. Horváth K, Liu S, Plunkett K. A daytime nap facilitates generalization of word meanings in young toddlers. Sleep. 2016;39(1):203-7. doi: 10.5665/sleep.5348.

11. Schapiro AC, McDevitt EA, Chen L, Norman KA, Mednick SC, Rogers TT. Sleep benefits memory for semantic category structure while preserving exemplar-specific information. Scientific reports. 2017;7(1):14869. doi: 10.1038/s41598-017-12884-5.
12. Hassabis D, Kumaran D, Vann SD, Maguire EA. Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(5):1726-31. doi: 10.1073/pnas.0610561104.
13. Gupta AS, van der Meer MA, Touretzky DS, Redish AD. Hippocampal replay is not a simple function of experience. Neuron. 2010;65(5):695-705. doi: 10.1016/j.neuron.2010.01.034.
14. Stella F, Baracskay P, O'Neill J, Csicsvari J. Hippocampal reactivation of random trajectories resembling brownian diffusion. Neuron. 2019;102(2):450-61.e7. doi: 10.1016/j.neuron.2019.01.052.
15. Deuker L, Olligs J, Fell J, Kranz TA, Mormann F, Montag C, et al. Memory consolidation by replay of stimulus-specific neural activity. The Journal of Neuroscience. 2013;33(49):19373-83. doi: 10.1523/jneurosci.0414-13.2013.
16. Ji D, Wilson MA. Coordinated memory replay in the visual cortex and hippocampus during sleep. Nature neuroscience. 2007;10(1):100-7. doi: 10.1038/nn1825.
17. de Voogd LD, Fernández G, Hermans EJ. Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. NeuroImage. 2016;134:563-72. doi: 10.1016/j.neuroimage.2016.04.026.
18. Tambini A, Ketz N, Davachi L. Enhanced brain correlations during rest are related to memory for recent experiences. Neuron. 2010;65(2):280-90. doi: 10.1016/j.neuron.2010.01.001.
19. Staresina BP, Alink A, Kriegeskorte N, Henson RN. Awake reactivation predicts memory in humans. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(52):21159-64. doi: 10.1073/pnas.1311989110.
20. Girardeau G, Inema I, Buzsáki G. Reactivations of emotional memory in the hippocampus–amygdala system during sleep. Nature neuroscience. 2017;20(11):1634. doi: 10.1038/nn.4637.
21. Eichenlaub J-B, Jarosiewicz B, Saab J, Franco B, Kelemen J, Halgren E, et al. Replay of learned neural firing sequences during rest in human motor cortex. Cell reports. 2020;31(5):107581. doi: 10.1016/j.celrep.2020.107581.
22. Peyrache A, Khamassi M, Benchenane K, Wiener SI, Battaglia FP. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nature neuroscience. 2009;12(7):919-26. doi: 10.1038/nn.2337.
23. Manns JR, Hopkins RO, Squire LR. Semantic memory and the human hippocampus. Neuron. 2003;38(1):127-33. doi: 10.1016/S0896-6273(03)00146-6.
24. Dudai Y. The restless engram: consolidations never end. Annual review of neuroscience. 2012;35:227-47. doi: 10.1146/annurev-neuro-062111-150500.
25. Giri B, Miyawaki H, Mizuseki K, Cheng S, Diba K. Hippocampal reactivation extends for several hours following novel experience. The Journal of Neuroscience. 2019;39(5):866-75. doi: 10.1523/JNEUROSCI.1950-18.2018
26. Drosopoulos S, Windau E, Wagner U, Born J. Sleep enforces the temporal order in memory. PloS one. 2007;2(4):e376. doi: 10.1371/journal.pone.0000376.
27. Kuriyama K, Stickgold R, Walker MP. Sleep-dependent learning and motor-skill complexity. Learning & memory. 2004;11(6):705-13. doi: 10.1101/lm.76304.
28. McDevitt EA, Duggan KA, Mednick SC. REM sleep rescues learning from interference. Neurobiology of learning and memory. 2015;122:51-62. doi: 10.1016/j.nlm.2014.11.015.
29. González OC, Sokolov Y, Krishnan GP, Delanois JE, Bazhenov M. Can sleep protect memories from catastrophic forgetting? eLife. 2020;9:e51005. doi: 10.7554/eLife.51005.
30. Kemker R, Kanan C. Fearnet: Brain-inspired model for incremental learning. arXiv preprint arXiv:171110563. 2017.
31. van de Ven GM, Siegelmann HT, Tolias AS. Brain-inspired replay for continual learning with artificial neural networks. Nature communications. 2020;11(1):4069. doi: 10.1038/s41467-020-17866-2.

32. 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(27):10005-14. doi: 10.1523/JNEUROSCI.5023-14.2015.

33. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology. 2014;10(11):e1003915. doi: 10.1371/journal.pcbi.1003915.

34. He K, Zhang X, Ren S, Sun J, editors. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; 2015.

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

36. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: Which artificial neural network for object recognition is most brain-Like? bioRxiv. 2018:407007. doi: 10.1101/407007.

37. Posner MI, Keele SW. On the genesis of abstract ideas. Journal of experimental psychology. 1968;77(3):353. doi: 10.1037/h0025953.

38. Donald H, Joseph C, Don C, David G, Steven S. Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. Journal of Experimental Psychology. 1973;101(1):116-22. doi: 10.1037/h0035772.

39. Reed SK. Pattern recognition and categorization. Cognitive Psychology. 1972;3(3):382-407. doi: 10.1016/0010-0285(72)90014-X.

40. Bowman CR, Iwashita T, Zeithamova D. Tracking prototype and exemplar representations in the brain across learning. eLife. 2020;9. doi: 10.7554/eLife.59360.

41. Mack ML, Love BC, Preston AR. Building concepts one episode at a time: The hippocampus and concept formation. Neuroscience letters. 2018;680:31-8. doi: 10.1016/j.neulet.2017.07.061.

42. Hassabis D, Kumaran D, Maguire EA. Using imagination to understand the neural basis of episodic memory. The Journal of Neuroscience. 2007;27(52):14365-74. doi: 10.1523/JNEUROSCI.4549-07.2007.

43. Mack ML, Preston AR, Love BC. Ventromedial prefrontal cortex compression during concept learning. Nature communications. 2020;11(1):46. doi: 10.1038/s41467-019-13930-8.

44. Giusberti F, Cornoldi C, De Beni R, Massironi M. Differences in vividness ratings of perceived and imagined patterns. British Journal of Psychology. 1992;83(4):533-47. doi: 10.1111/j.2044-8295.1992.tb02457.x.

45. Lee SH, Kravitz DJ, Baker CI. Disentangling visual imagery and perception of real-world objects. NeuroImage. 2012;59(4):4064-73. doi: 10.1016/j.neuroimage.2011.10.055.

46. Winocur G, Moscovitch M. Memory transformation and systems consolidation. Journal of the International Neuropsychological Society : JINS. 2011;17(5):766-80. doi: 10.1017/S1355617711000683.

47. Sweegers CCG, Talamini LM. Generalization from episodic memories across time: A route for semantic knowledge acquisition. Cortex; a journal devoted to the study of the nervous system and behavior. 2014;59:49-61. doi: 10.1016/j.cortex.2014.07.006.

48. Love BC, Medin DL. SUSTAIN: A model of human category learning. Aaai/iaai. 1998:671-6.

49. Rothschild G, Eban E, Frank LM. A cortical-hippocampal-cortical loop of information processing during memory consolidation. Nature neuroscience. 2017;20(2):251-9. doi: 10.1038/nn.4457.

50. Poe GR, Nitz DA, McNaughton BL, Barnes CA. Experience-dependent phase-reversal of hippocampal neuron firing during REM sleep. Brain research. 2000;855(1):176-80. doi: 10.1016/S0006-8993(99)02310-0.

51. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nature neuroscience. 2019;22(11):1761-70. doi: 10.1038/s41593-019-0520-2.

52. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. 2009:248-55.
53. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014;15(1):1929-58.
54. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision. 2015;115(3):211-52. doi: 10.1007/s11263-015-0816-y.
55. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
56. Remy P. Keract: A library for visualizing activations and gradients. GitHub repository. 2020.
57. Fan Y, Tian F, Qin T, Li X-Y, Liu T-Y. Learning to teach. arXiv preprint arXiv:180503643. 2018.

Supplementary table S1: 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