’Reusers’ and ’Unlearners’ display distinct effects of forgetting on reversal learning in mice and artificial neuronal networks

Jonas Elpelt (elpelt@fias.uni-frankfurt.de)
Department of Computer Science, Goethe University Frankfurt
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Jens-Bastian Eppler (eppler@fias.uni-frankfurt.de)
Department of Computer Science, Goethe University Frankfurt
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Johannes P.-H. Seiler (johseile@uni-mainz.de)
Institute of Physiology, University Medical Center
Johannes Gutenberg University Mainz
Hanns-Dieter-Hüsch-Weg 19, 55131 Mainz

Simon Rumpel (sirumpel@uni-mainz.de)
Institute of Physiology, University Medical Center
Johannes Gutenberg University Mainz
Hanns-Dieter-Hüsch-Weg 19, 55131 Mainz

Matthias Kaschube (kaschube@fias.uni-frankfurt.de)
Department of Computer Science, Goethe University Frankfurt
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Abstract

Previous research has indicated that prior learning can be both advantageous or disadvantageous for learning related tasks. Moreover, the speed of learning a related task might be mediated by the extent of forgetting of the original task. Here, we seek to explore the role of forgetting initially learned task representations on reversal learning behavior in both mice and artificial recurrent neural networks. We trained mice to discriminate two auditory stimuli in a go/no-go paradigm. After learning, they had a pause of 2 or 16-days. In general, a shorter pause resulted in better memory retention and faster adaptation during reversal learning with reversed conditions. However, some animals did not benefit from initial learning, suggesting no reuse of initial representations. Similar patterns were observed in artificial neural networks during reversal learning, showing both beneficial reusing and disadvantageous unlearning of previously learned network configurations. Our findings shed light on the use of initial representations during reversal learning and could provide insights into cognitive flexibility in both biological and artificial neural networks.

Keywords: forgetting; reversal learning; recurrent neural networks; adaptation; cognitive flexibility; unlearning

Introduction

Neural networks need to continuously adapt their representations to survive in a constantly changing environment. This flexible adaptation of entrained representations can be mediated by unlearning and forgetting (Guskjolen & Cembrowski, 2023). One possible factor for passive forgetting could be the spontaneous remodeling of neuronal circuitry (Davis & Zhong, 2017). In this study, we examine how forgetting impacts performance in a reversal learning task, in which we trained animals to perform a go/no-go discrimination paradigm and then invert the initial reward contingencies. Acknowledging that previous studies have presented conflicting views on the effects of forgetting during such learning schemes, we hypothesize two potential scenarios (Fig. 1).

Hypothesis I ('Reusing'): Reusing memory from a previously learned task could be beneficial for the reversal task, because parts of the structure of the initially learned task could be reused (Gonzalez, Behrend, & Bitterman, 1967; Woodworth & Thorndike, 1901; Day & Goldstone, 2012; Bransford & Schwartz, 1999). In this case higher forgetting of the initial task would lead to slower learning of the reversal task.

Hypothesis II ('Unlearning'): Reusing previously learned representations could have no advantage for reversal learning, because representations of the new task could interfere with previously learned associations (Bouton, Nelson, & Rosas, 1999; Luchins, 1942; Wixted, 2004; Postman, 1971). In this case higher forgetting of the initial task could even accelerate learning of the reversal task.

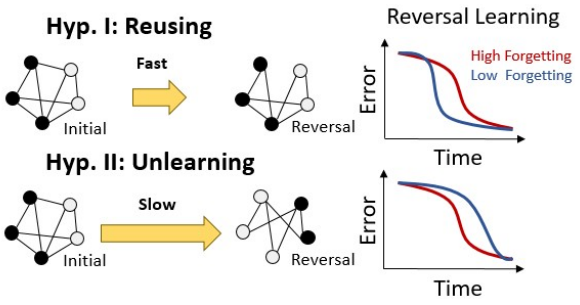

Figure 1: **Illustration of potential scenarios** Hypothesis I: Initial learning speeds up the reversal learning process. In this case forgetting impairs reversal learning as it prohibits reusing previously learned structures. Hypothesis II: Initial learning slows down the reversal learning process. In this case forgetting aids reversal learning as it prevents interference with previously learned structures and supports unlearning.

Methods

Experimental Design We trained mice in an operant learning task to discriminate two auditory stimuli in a go/no-go paradigm. 16 wildtype C57BL/6J mice were trained on a go/no-go discrimination task with two pulsed auditory cues. After an initial learning period of three weeks, mice were separated pseudorandomly into two equally-sized groups that had a passive pause interval of either 2 or 16 days. After the pause, we probed memory retention by testing the performance of the mice on the initial task, followed by a reversal learning task with inverted contingencies for both stimuli.

Model specifications We employed a single-layer recurrent neural network (RNN) with 2-dimensional input over t = 40 time points to perform a binary classification, equivalent to the previously described experiment. The RNN consists of n = 50 hidden units with a non-linear activation function (hyperbolic tangent). The network was trained with backpropagation through time (stochastic gradient descent, learning rate = 0.01) (Werbos, 1990) and a dropout rate d = 0.3. Mean squared error (MSE) was used as loss function. The recurrent weights W were initialized with random normal distribution with specified mean (between 0 and -0.6) and standard deviation (between 0.1 and 0.9), which have been kept constrained throughout training. Forgetting is implemented by randomly shuffling a fraction of W after initial learning.

To distinguish 'Reusers' from 'Unlearners' in the RNN we fitted sigmoidal curves to the reversal learning curves and computed rank correlations between the delay in learning and the proportion of shuffled weights. We expected 'Reusers' to have a positive correlation (Hypothesis I) and 'Unlearners' to have a small or negative correlation (Hypothesis II) (Fig. 1).

To illustrate the difference between representations during reversal learning we computed 'Weight Change Norm' (WC) and 'Representation Alignment' (RA) in the RNN (Liu et al., 2023). WC is computed by: $\| \Delta W \| := \| W^{(r)} - W^{(i)} \|$,

where $W^{(i)}$ (resp. $W^{(r)}$) are the recurrent weights after initial (resp. reversal) training. RA is computed by: $RA(R^{(r)}, R^{(i)}) := \frac{Tr(R^{(r)}, R^{(i)})}{\|R^{(r)}\|\|R^{(i)}}$, where $R^{(i)}$ (resp. $R^{(r)}$) is the representational similarity matrix after initial (resp. reversal) training.

Results

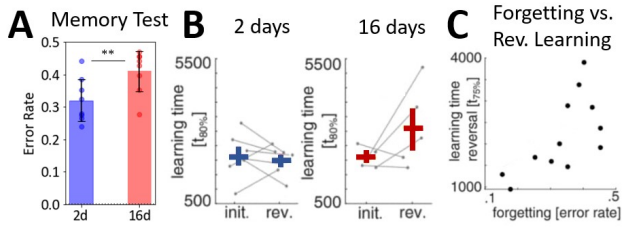

Figure 2: **Forgetting depends on pause and has impact on reversal learning in mice** A) Forgetting in memory test is higher after longer pause between learning and testing B) Forgetting has different impact on mice after a pause of 2 days resp. 16 days C) Learning time of reversal task is positively correlated with forgetting.

Mice display different performances in reversal learning after pause

When mice underwent reversal learning after a pause of either 2 days or 16 days from the initial learning the length of the pause was positively associated with the error rate in a memory test (Fig. 2). Therefore a longer pause after initial training seemed to lead to higher forgetting of the original task. Moreover, forgetting was positively correlated with reversal learning time, showing a pronounced delay during learning of the reversed contingency task with more forgetting. However, not all animals did show an advantage of learning the initial task when comparing learning dynamics between initial and reversal tasks.

Reversal learning performance depends on initially learned network configuration and level of forgetting

In the experiment we observed both positive and negative effects of initial learning on the performance during the reversal task at an individual level. When learning curves were averaged over multiple animals a delay in reversal learning was visible for a longer pause between initial and reversal learning. RNNs captured this behavior (Fig. 4). Interestingly, the effect of forgetting on reversal learning depended on different initially learned configurations. The majority of trained RNNs were 'Reusers', which profited from initial learning of the opposite task. However, a subset of networks were 'Unlearners', which were not capable of using previously learned representations to speed up the reversal learning process (Fig. 3). 'Reusers' showed higher influence of forgetting on the representational alignment of their neuronal activity and weight changes within the recurrent layer after reversal learning. This indicated a profitable network configuration for switching stimulus contingencies in contrast to 'Unlearners'.

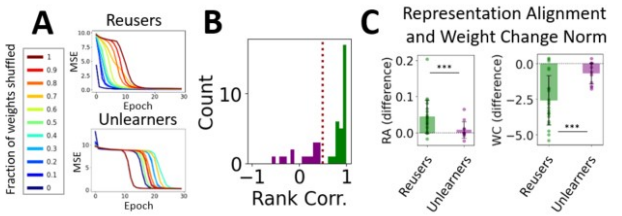

Figure 3: **Reversal Learning performance depends on individual network configurations after initial learning** A) Examples of networks with high and low rank correlation between forgetting and learning delay. B) Histogram of rank correlations (We distinguish between 'Reusers' with a rank correlation > 0.5 and 'Unlearners' with a rank correlation < 0.5.) C) 'Reusers' show higher differences between 0% and 100% shuffling of weights during reversal learning, suggesting a higher impact of forgetting on the capability to reuse initially learned representations.

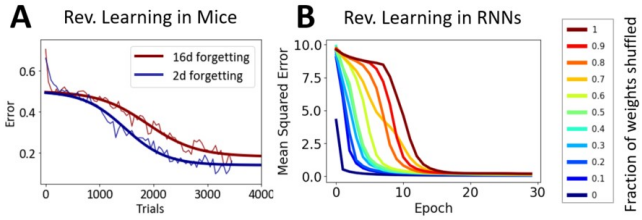

Figure 4: **Reversal learning performance depends on level of forgetting.** A) Average reversal learning curve for a cohort of mice tested after 16 days displays slower reversal learning than a cohort tested after 2 days. B) Reversal learning curves in the network model after shuffling different fractions of synaptic weights display slower reversal learning for higher forgetting.

Discussion

In conclusion, our study sheds light on the interplay of passive forgetting, neural representations and reversal learning: Delayed reversal learning observed in experimental data averaged over multiple animals can be successfully simulated by a constrained RNN with shuffled synaptic weights mimicking passive forgetting. However, we found evidence for both beneficial and disadvantageous forgetting depending on individual network configurations after initial learning. In future research we aim to characterize these individual differences in animals and RNNs and plan to acquire imaging data in mice to compare neural representation changes *in-vivo*. Overall, our findings provide valuable insights into memory dynamics and may have implications for understanding cognitive flexibility and adaptation in both biological and artificial neural systems on a network level.

Acknowledgments

This research has received funding from DFG-Project 450290950 'The Neurobiology of Forgetting' (JE, JPHS, SR, MK).

References

Bouton, M. E., Nelson, J. B., & Rosas, J. M. (1999). Stimulus generalization, context change, and forgetting. *Psychol. Bull.*, *125*(2), 171–186.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61.

Davis, R. L., & Zhong, Y. (2017). The biology of forgetting—a perspective. *Neuron*, *95*(3), 490–503.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153–176.

Gonzalez, R. C., Behrend, E. R., & Bitterman, M. E. (1967). Reversal learning and forgetting in bird and fish. *Science*, *158*(3800), 519–521.

Guskjolen, A., & Cembrowski, M. S. (2023). Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular Psychiatry*, *28*(8), 3207–3219.

Liu, Y. H., Baratin, A., Cornford, J., Mihalas, S., Shea-Brown, E., & Lajoie, G. (2023). How connectivity structure shapes rich and lazy learning in neural circuits. *arXiv*.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological Monographs*, *54*(6), i–95.

Postman, L. (1971). Transfer, interference and forgetting. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth and Schlosberg's experimental psychology* (Third ed., pp. 1019–1132). New York: Holt, Rinehart and Winston.

Werbos, P. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550-1560.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*(1), 235–269.

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (i). *Psychological Review*, *8*(3), 247–261.