—PhD Research Topic—
Active Forgetting Across Brains and Models: Neuro-Computational Insights

Research axis of the 3IA: Axis 3 - AI for Computational Biology and Bio-inspired AI
Supervisor (3IA Chair): Emanuele Natale, Sophia Antipolis Laboratory for Computer Science, Signals and Systems (I3S), Sophia Antipolis
Co-supervisor (co-encadrant): Bianca Silva, Institute of Molecular and Cellular Pharmacology (IPMC), Sophia Antipolis
Hosting lab: I3S & INRIA UniCA

Apply by sending an email directly to the supervisor:
emanuele.natale@univ-cotedazur.fr

Primary discipline: Machine Learning
Secondary discipline: Neuroscience

Project Summary

This project proposes to explore how the brain and machines unlearn. In everyday life, disassociating memories is essential—letting us move on from fears, mistakes, or outdated beliefs. However, how memory systems in our brains achieve this, remains unclear. Similarly, forgetting is a challenge for artificial intelligence: once a machine learns something, it's hard to have it forget. This has unwanted implications when machines learn something wrong, private, copyrighted, or biased. By studying brain data recordings and building computational models that mimic real populations of neurons, the project aims to uncover active unlearning: how the brain learns to dissociate memories. Finally, it proposes to use this information to come up with novel strategies to make machines unlearn better, more efficiently, and more safely. The potential impact of this project is twofold: i) understanding how the brain unlearns, may help us design new strategies against mental health conditions in which unlearning is impaired, such as post-traumatic stress disorder or addiction. Simultaneously, this project has the potential to ii) improve AI by introducing efficient and direct unlearning, thus enabling better handling of fake information, harmful associations or private data.

Scientific Description

This PhD project bridges computational neuroscience and machine learning to study the mechanisms of active forgetting—or unlearning—through the lens of both biological and artificial systems. Unlearning is crucial for biological organisms to adapt and remain flexible in dynamic environments, as well as for machines to optimize output integrity by shedding outdated or harmful associations. In this project we will draw analogies between memory dynamics in rodent brains and challenges in machine unlearning, particularly in foundation models such as Large Language Models. Using experimental data from rodent studies—obtained by electrophysiology, fiber photometry, and calcium imaging in vivo recordings—the project will build a biologically grounded SNN model of memory de-association. The ultimate goal is to both validate this model through experiments and apply its insights to enhance auditability and ethical deletion protocols in artificial systems.

Scientific aim

This project aims to investigate the neuro-computational mechanisms of unlearning, drawing analogies between artificial intelligence (AI) and biological neural systems. In particular, it first seeks to unravel how the brain unlearns previously encoded information through the analysis of neural recordings of large neuronal populations in key memory areas. Specifically, we aim to build a general model of unlearning that combines diverse experimental data and is testable across multiple unlearning scenarios. For this we plan to apply for the first time Spiking Neural Networks (SNNs) to the **modeling of unlearning.** SNNs have recently shown enhanced efficiency in modeling learning and synaptic plasticity [1-3] in biologically plausible architectures, but have not been used to unravel the computational principles of memory unlearning. By doing this, it aims at assessing whether memory unlearning consists of memory **erasure** (disruption of an existing memory trace, referred as *reconsolidation update* in the field) or memory **sidelearning** (creation of a new memory trace that inhibits but does not delete the original one, referred as *extinction learning* in the field), a currently unresolved question in the field [4].

Addressing this issue would substantially contribute to our current understanding of how the brain unlearns, a fundamental function whose impairments are at the core of a number of brain disorders, including PTSD, chronic pain, and anxiety disorders. Simultaneously, unraveling the coding rules of unlearning has the potential to help inspire new strategies for enabling efficient unlearning of target information into LLMs.

Scientific context

Large Language Models (LLMs) are trained on massive text datasets—often in the order of terabytes—making it almost impossible to filter out undesirable or outdated information. When wrong, private, or copyrighted information is used for training, it often compromises the models with undesired associations that must be subsequently severed. Re-training the model may not be always viable, and it easily becomes necessary to prevent and remove particular associations within the model, without disrupting its integrity. Inevitably, one question arises: how can we make LLMs forget? This challenge is referred to as **Machine Unlearning**. The unlearning literature for machine learning models can roughly be categorized into the following: exact unlearning, "unlearning" via differential privacy and empirical unlearning [5].

The problem is not dissimilar to that faced by biological organisms, which have evolved efficient strategies to update outdated or irrelevant information. In neuroscience research, a large body of literature has investigated the neural mechanisms that govern such memory unlearning processes.

In rodent models, a classical example of unlearning, in which animals learn to de-associate a previously encoded associative memory, is the paradigm of **fear memory extinction**. In this paradigm repeated exposure to a non-reinforced conditioned stimulus—such as a tone without the expected shock—leads to the progressive suppression of a previously encoded fear memory. This paradigm has been widely used to investigate the neural mechanisms at the basis of the dissociation of previously linked stimuli. As such, it has generated a vast body of data and detailed characterization of the associated memory circuits and their dynamics. This vast body of literature has led to the notion that memories depend on distributed neuronal assemblies known as **engrams**, that are later modified for memory erasure.

However, the nature of these modifications is still unclear. The engram networking rules mediating memory formation have been recently described by SNNs with great promise, yet their applicability to unlearning remains to be explored. Here we propose to build a data-driven SNN model based on the extensive experimental data from memory extinction paradigms to unpack the neuronal mechanisms of unlearning. Among these mechanisms, one central question is whether unlearning is mediated by the disruption of the original engram (memory erasure) or through the formation of a new memory trace that competes with the original while rendering it **silent** (memory sidelearning). Understanding the variables at play could help address a critical issue associated with silent engrams: while erasure leaves no trace, sidelearning preserves memories below the behavioural threshold, which could still be reactivated under certain conditions. In a parallel fashion, machine learning models may be unable to fully and safely erase complex/distributed undesired information due to the implication it would have on their overall integrity, leaving "sidelearning" as the only viable option. This suggests a compelling link between the neuroscientific concept of silent engrams and the **auditability problem** in AI: how can one verify that a model no longer retains a particular piece of information after a deletion request?

Scientific methodology

The methodology of this project is structured around three main aims, each designed to address a core aspect:

Aim 1: *Build a dataset capturing the neural dynamics of fear memory extinction*

1.1. Analysis of the state of the art of spiking neural networks, particularly their use in learning and memory paradigms.

1.2. Gather and curate neural activity recording datasets investigating fear memory extinction in rodents [5-11].

1.3. Develop a consistent framework to integrate the heterogeneous data (e.g., fiber photometry, calcium imaging, or electrophysiology recordings) identified in step 1.2. TheVirtualBrain framework [12] will assist in effectively managing the integration process.

Aim 2: *Model the computational de-association of memory traces*

2.1. Use abstractions and frameworks to represent complex neuronal assemblies with a small number of computational elements [1-3].

2.2. Design and develop SNN model to simulate memory decoupling (i.e., unlearning of associations), distinguishing between memory sidelearning and memory erasure.

2.3. Incorporate physics-informed neural networks [13] (PINNs) and Graph Neural Networks [14[(GNNs) to embed biological constraints, improving model generalization under sparse data conditions. This step will build upon the integrated multimodal data layer developed earlier and will aim to find the parameters in the SNN that best capture the dynamics of excitatory and inhibitory populations during the unlearning process.

2.4. Design targeted neural recording experiments to fill potential gaps in the available datasets (specific cell types, brain areas, behavioral epochs). Experimental recordings will be performed by the Silva team (co-supervisor) that routinely runs in vivo calcium imaging during fear extinction behavioral paradigms in mice.

Aim 3: *Validate the model and generate testable predictions*

3.1. Evaluate key variables within the model to investigate overwrite vs sidelearning in memory unlearning.

3.2. Use the data-driven model to generate hypotheses on the nature of unlearning and the conditions under which memories persist after fear extinction.
3.3. Collaborate with experimentalists (Silva team) to design and test these predictions experimentally in rodents.

Interdisciplinary Impact

The project is grounded in a strong experimental foundation, generating testable hypotheses for validation in rodents. At the same time, it contributes to basic research on LLMs by integrating biological insights into machine learning. By developing a biologically informed computational model of active forgetting, the project seeks to clarify when and how memory erasure or sidelearning strategies are preferable—both in biological and artificial systems. Crucially, it aims to inform ethical and efficient machine unlearning strategies by providing a theoretical framework linking the neuroscientific concept of **silent engrams** to the **auditability problem** in AI.

Originality

No comparable effort has yet been devoted to modeling unlearning or de-association processes in SNNs, despite the fact that they have been used with great success to model learning in mice [1-3]. This project aims to address this gap. Moreover, the originality of the project stands on the fact that activity recording data are collected and integrated in the model from multiple experimental sources, in the hope to exploit the full power of computational modelling to span over different orders of magnitude in both time and space—something crucial in neuroscience but not easily feasible for experimental neuroscience or by reducing the modeling to in-lab produced data.
Additionally, the project explores the use of Physics-Informed Neural Networks (PINNs) to guide the parameter setting of the SNNs based on experimental data. This allows for a more unbiased and data-driven approach to model calibration, overcoming limitations of previous models that often relied on hand-tuned or assumption-heavy parameterization.

Workplan

The work plan is structured in three phases corresponding to the methodological aims:

- **Phase 1 (Months 1–12):** Literature review, data collection from rodent studies, and development of the integration framework.
  Given the plethora of data and interest in the field to investigate on the fear memory circuit, there is negligible risk that we will have an irremediable gap in the data that tampers our possibilities to feed the model for our purposes.
- **Phase 2 (Months 13–24):** Design and implementation of SNN models, incorporation of biological constraints using PINNs and GNNs.
  SNN already found success in modelling learning circuits, an adjacent task to unlearning [1-3]. Meanwhile, GNNs have shown promising results in connectomics analyses [14-21], though their application to neuro-modelling remains largely unexplored. Similarly, while PINNs are widely used in other scientific domains [13,22-24], their integration into neuro-modelling is still in its early stages.
- **Phase 3 (Months 25–36):** Model validation, hypothesis generation, and experimental collaboration to test predictions in rodents.

The project builds upon available experimental paradigms and computational tools, ensuring the feasibility of each stage within a typical PhD timeline.

References

(1) Tomé, D. F.; Zhang, Y.; Aida, T.; Mosto, O.; Lu, Y.; Chen, M.; Sadeh, S.; Roy, D. S.;

Clopath, C. Dynamic and Selective Engrams Emerge with Memory Consolidation. *Nat. Neurosci.* **2024**, 1–12. https://doi.org/10.1038/s41593-023-01551-w.

(2) Tomé, D. F.; Sadeh, S.; Clopath, C. Coordinated Hippocampal-Thalamic-Cortical Communication Crucial for Engram Dynamics underneath Systems Consolidation. *Nat. Commun.* **2022**, *13* (1), 840. https://doi.org/10.1038/s41467-022-28339-z.

(3) Zenke, F.; Agnes, E. J.; Gerstner, W. Diverse Synaptic Plasticity Mechanisms Orchestrated to Form and Retrieve Memories in Spiking Neural Networks. *Nat. Commun.* **2015**, *6* (1), 6922. https://doi.org/10.1038/ncomms7922.

(4) Kida, S. Chapter Six - Memory Reconsolidation Versus Extinction. In *Memory Reconsolidation*; Alberini, C. M., Ed.; Academic Press: San Diego, 2013; pp 119–137. https://doi.org/10.1016/B978-0-12-386892-3.00006-8.

(5) Grewe, B. F.; Gründemann, J.; Kitch, L. J.; Lecoq, J. A.; Parker, J. G.; Marshall, J. D.; Larkin, M. C.; Jercog, P. E.; Grenier, F.; Li, J. Z.; Lüthi, A.; Schnitzer, M. J. Neural Ensemble Dynamics Underlying a Long-Term Associative Memory. *Nature* **2017**, *543* (7647), 670–675. https://doi.org/10.1038/nature21682.

(6) Favila, N.; Marsico, J. C.; Escribano, B.; Pacheco, C. M.; Bitterman, Y.; Gründemann, J.; Lüthi, A.; Krabbe, S. Heterogeneous Plasticity of Amygdala Interneurons in Associative Learning and Extinction. bioRxiv September 29, 2024, p 2024.09.29.612271. https://doi.org/10.1101/2024.09.29.612271.

(7) Taylor, J. A.; Hasegawa, M.; Benoit, C. M.; Freire, J. A.; Theodore, M.; Ganea, D. A.; Innocenti, S. M.; Lu, T.; Gründemann, J. Single Cell Plasticity and Population Coding Stability in Auditory Thalamus upon Associative Learning. *Nat. Commun.* **2021**, *12* (1), 2438. https://doi.org/10.1038/s41467-021-22421-8.

(8) Senn, V.; Wolff, S. B. E.; Herry, C.; Grenier, F.; Ehrlich, I.; Gründemann, J.; Fadok, J. P.; Müller, C.; Letzkus, J. J.; Lüthi, A. Long-Range Connectivity Defines Behavioral Specificity of Amygdala Neurons. *Neuron* **2014**, *81* (2), 428–437. https://doi.org/10.1016/j.neuron.2013.11.006.

(9) Venkataraman, A.; Brody, N.; Reddi, P.; Guo, J.; Gordon Rainnie, D.; Dias, B. G. Modulation of Fear Generalization by the Zona Incerta. *Proc. Natl. Acad. Sci.* **2019**, *116* (18), 9072–9077. https://doi.org/10.1073/pnas.1820541116.

(10) Silva, B. A.; Astori, S.; Burns, A. M.; Heiser, H.; van den Heuvel, L.; Santoni, G.; Martinez-Reza, M. F.; Sandi, C.; Gräff, J. A Thalamo-Amygdalar Circuit Underlying the Extinction of Remote Fear Memories. *Nat. Neurosci.* **2021**, *24* (7), 964–974. https://doi.org/10.1038/s41593-021-00856-y.

(11) Hagihara, K. M.; Bukalo, O.; Zeller, M.; Aksoy-Aksel, A.; Karalis, N.; Limoges, A.; Rigg, T.; Campbell, T.; Mendez, A.; Weinholtz, C.; Mahn, M.; Zweifel, L. S.; Palmiter, R. D.; Ehrlich, I.; Lüthi, A.; Holmes, A. Intercalated Amygdala Clusters Orchestrate a Switch in Fear State. *Nature* **2021**, *594* (7863), 403–407. https://doi.org/10.1038/s41586-021-03593-1.

(12) Ritter, P.; Schirner, M.; McIntosh, A. R.; Jirsa, V. K. The Virtual Brain Integrates Computational Modeling and Multimodal Neuroimaging. *Brain Connect.* **2013**, *3* (2), 121–145. https://doi.org/10.1089/brain.2012.0120.

(13) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.* **2019**, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

(14) Mohammadi, H.; Karwowski, W. Graph Neural Networks in Brain Connectivity Studies: Methods, Challenges, and Future Directions. *Brain Sci.* **2025**, *15* (1), 17. https://doi.org/10.3390/brainsci15010017.

(15) Zheng, K.; Yu, S.; Chen, B. CI-GNN: A Granger Causality-Inspired Graph Neural Network for Interpretable Brain Network-Based Psychiatric Diagnosis. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2024**, *172*, 106147.

https://doi.org/10.1016/j.neunet.2024.106147.

(16) Wein, S.; Malloni, W. M.; Tomé, A. M.; Frank, S. M.; Henze, G.-I.; Wüst, S.; Greenlee, M. W.; Lang, E. W. A Graph Neural Network Framework for Causal Inference in Brain Networks. *Sci. Rep.* **2021**, *11* (1), 8061. https://doi.org/10.1038/s41598-021-87411-8.

(17) Zhang, H.; Song, R.; Wang, L.; Zhang, L.; Wang, D.; Wang, C.; Zhang, W. Classification of Brain Disorders in Rs-fMRI via Local-to-Global Graph Neural Networks. *IEEE Trans. Med. Imaging* **2023**, *42* (2), 444–455. https://doi.org/10.1109/TMI.2022.3219260.

(18) Bessadok, A.; Mahjoub, M. A.; Rekik, I. Graph Neural Networks in Network Neuroscience. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45* (5), 5833–5848. https://doi.org/10.1109/TPAMI.2022.3209686.

(19) Li, Z.; Hwang, K.; Li, K.; Wu, J.; Ji, T. Graph-Generative Neural Network for EEG-Based Epileptic Seizure Detection via Discovery of Dynamic Brain Functional Connectivity. *Sci. Rep.* **2022**, *12* (1), 18998. https://doi.org/10.1038/s41598-022-23656-1.

(20) Yang, Y.; Ye, C.; Cai, G.; Song, K.; Zhang, J.; Xiang, Y.; Ma, T. Hypercomplex Graph Neural Network: Towards Deep Intersection of Multi-Modal Brain Networks. *IEEE J. Biomed. Health Inform.* **2024**, 1–13. https://doi.org/10.1109/JBHI.2024.3490664.

(21) Wein, S.; Malloni, W. M.; Tomé, A. M.; Frank, S. M.; Henze, G.-I.; Wüst, S.; Greenlee, M. W.; Lang, E. W. A Graph Neural Network Framework for Causal Inference in Brain Networks. *Sci. Rep.* **2021**, *11* (1), 8061. https://doi.org/10.1038/s41598-021-87411-8.

(22) Cai, S.; Mao, Z.; Wang, Z.; Yin, M.; Karniadakis, G. E. Physics-Informed Neural Networks (PINNs) for Fluid Mechanics: A Review. *Acta Mech. Sin.* **2021**, *37* (12), 1727–1738. https://doi.org/10.1007/s10409-021-01148-1.

(23) Cai, S.; Wang, Z.; Wang, S.; Perdikaris, P.; Karniadakis, G. E. Physics-Informed Neural Networks for Heat Transfer Problems. *J. Heat Transf.* **2021**, *143* (060801). https://doi.org/10.1115/1.4050542.

(24) Cuomo, S.; Di Cola, V. S.; Giampaolo, F.; Rozza, G.; Raissi, M.; Piccialli, F. Scientific Machine Learning Through Physics–Informed Neural Networks: Where We Are and What's Next. *J. Sci. Comput.* **2022**, *92* (3), 88. https://doi.org/10.1007/s10915-022-01939-z.