School of Biochemistry and Immunology

Trinity College Dublin

Engrams reconsidered: three windows on memory traces

by

Fionn M. O'Sullivan

A Dissertation

Presented to the University of Dublin, Trinity College

in fulfilment of the requirement for the degree of Research Masters

Master of Science in Biochemistry

Supervisor: Dr. Tomás Ryan

2025

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

No artificial intelligence software was used at any point in the production of this work, either in the process of finding literature, generating summaries, or the writing of the text itself.

Name: Fionn O'Sullivan

Date: 2024

Abstract

At the heart of this thesis is a simple problem—we remember, and yet no memory has ever been found. Does this problem derive from the experimental limitations of neuroscience, or is there a problem with our thinking? In other words, is the problem difficult because it is a hard scientific problem, or because we have assumed too much of the ideas we take for granted? This thesis seeks to improve how we think about the concepts of "memory" and "information processing and storage" in biological systems, through pooling together established literature and emerging research in neuroscience, philosophy, and machine learning. Through a combination of criticism, cross-fertilisation and elaboration of new ideas, it synthesises and presents more critical ways to ask questions about memory for a variety of audiences (neuroscience, philosophy, computer science/machine learning, and the cognitive sciences generally) with differing interests and objectives, and which each field can develop further. Chapter 1 provides a critical assessment and taxonomy of memory trace concepts and underscores the need for neuroscience to expand beyond characterising putative memory traces vehicles. Chapter 2 showcases what engagement with memory neuroscience has to offer neuroscience-inspired artificial intelligence, such as a shift away from spike-centric modelling. Finally, Chapter 3 outlines the shortcomings of philosophical accounts of representation for focusing only on spiking activity, and offers a novel defence of why existing concepts of computation and information processing cannot be applied to living systems.

Acknowledgements

I would like to thank first and foremost Prof. Tomás Ryan for his guidance, feedback, and critical appraisals throughout this research project. In particular, I am especially grateful for his support of blue skies research not only in the domain of fundamental empirical science but also in the world of scientific ideas.

I would like to thank all members of the Ryan Lab for allowing a non-experimentalist to live within their world. Thank you Dr. Clara Ortega-de San Luis, Dr. James O'Leary, Dr. Aaron Douglas, Dr. Livia Autore, Esteban Urieta, Dr. Andrea Munez Zamora, Dr. Paul Conway, Dr. Erika Stobart, Louisa Zielke, and other past and visiting researchers. I am immensely grateful for your practical feedback, insightful opinions, scientific advice and willingness to engage in philosophical discussions over the past two years.

I am grateful to faculty of the Trinity College Institute of Neuroscience for creating a general ethos that recognises the value of interdisciplinary research. An environment where scientists are willing to support projects that involve having their work and ideas scrutinised and where the gains are uncertain and hard to define in advance is worth cherishing.

I would like to thank Prof. Tomás Ryan and Dr. Francis Fallon for allowing me to participate in the regular meetings of the "Representations: Past, Present, Future" working group, and faculty and students at the Sanfe Fe Institute's 2023 "Intelligence & Representation" Complexity-GAINs international summer school.

I would also like to thank faculty at the School of Social Sciences and Philosophy, School of Linguistics, Speech and Communication Sciences at Trinity, and the School of Computer Science at University College Dublin for kindly allowing me to audit relevant modules early on in this research project.

I am grateful to Phelim O’Laoghaire, Dr. Henry Potter, Jane Cook, and Caitlin Mace for numerous discussions and excellent points of feedback on versions of this work over the past two years.

Finally, I would like to thank my family for their unwavering support, the Trinity disAbility Service, Student Counselling Service, and the other staff and departments at Trinity that have supported me throughout my studies.

This research project would not have been possible or taken the shape it has without these people.

Thank you.

"Nothing should be told about the locations in the brain of the faculties of the supreme spirit."

— Andreas Vesalius (as cited in Cobb, 2020)

This thesis is dedicated to the legacy of Śāntarakṣita (c.725–788).

Table of Contents

Introduction & Methodology

Understanding how and what is stored in the brain is one of the great unknowns in modern science. Given its usage in everyday parlance, it is surprising to learn that "memory" is a folk-psychological concept that does not yet have a clear biological explanation. Given its usage in everyday scientific parlance, it is also surprising to find that the more technically sounding "information storage" does not refer to a scientific theory of memory either. What do we mean when we say the brain "stores information"? What exactly are we looking for in the brain when we search for memory traces? Memory is an area of central importance in the cognitive sciences but at present remains theoretically under-researched (in comparison to consciousness studies, for example).

Despite decades of research on learning and memory, fundamental unknowns remain about how the brain acquires and holds knowledge. To date there exists no formal framework that rivals other constructs, like Shannon information, which was not developed for biological architectures. The absence of such a unique framework is felt across biology but most keenly in the relatively young field of neuroscience, where conceptual advances have not kept pace with methodological developments, resulting in a theoretical crisis that only deepens as the years progress. An ad-hoc hermeneutics of neural data has emerged, with neuroscientists often drawing uncritically on concepts from psychology, genetics, computer science and mathematics, which can carry confusing connotations.

This thesis interrogates the theoretical foundations and experimental methods that have been used to understand memory, providing neuroscientists and philosophers with different conceptions of and issues with what is meant by information storage in neuroscience. Although presented in three different chapters for different audiences, throughout I seek to clarify why theories based on finding von-Neuman architectures and the digital bits of Shannon information theory in the brain are misleading. I end by proposing an alternative biologically-inspired

account of computation that accommodates recent work on the systems neuroscience of memory.

General Epistemological Orientation

Despite being formally hosted by the School of Biochemistry and Immunology, this is a theoretical rather than experimental research masters thesis. Interdisciplinary research is much encouraged but equally fraught, demanding familiarity with expanded bodies of literature, all the more so given that neuroscience is not a coherent whole. Callard and Fitzgerald's (2015) *Rethinking interdisciplinarity across the social sciences and neurosciences* reassured me against the image of interdisciplinary research as inherently risky, but that rather not playing to disciplinary norms can be a valid and useful research strategy.

As an interdisciplinary thesis it may be variously categorised. Since no one descriptor adequately encapsulates this thesis as a whole, I offer several. A broadest domain this project sits within is "cognitive science," given this label includes neuroscience, philosophy and computer science. The next level of description would be that it sits at the intersection of philosophy of science and theoretical biology. With the move away from domain general philosophy of science to the special sciences, this thesis falls within the contemporary areas of philosophy of biology and philosophy of neuroscience, while not being a straight "history and philosophy of science" or "philosophy of mind" thesis. Additionally, while it may further be considered within "philosophy of memory"—contributing to the recent philosophy of memory subdivision of philosophy of science, I deal almost exclusively with recent systems neuroscience rather than the cognitive neuroscience and cognitive psychology of humans that is primarily discussed in this philosophical literature. Finally, because I am drawing from neuroscience to inform theorising, there is a mixture of neurophilosophy and philosophy of neuroscience.

I subscribe to a general principle popular in social science research (e.g. Friedrichs & Kratochwil, 2009) that researchers, regardless of whether we are doing empirical or theoretical work, should state clearly and publicly the purposes of our research as well as personal motivations. I have covered the purposes and general motivations in the introduction. In terms of other personal motivations, while this is not a Science and Technology Studies thesis, given that this project involved being a member of an experimental systems neuroscience laboratory, sharing an office and weekly laboratory presentations of experimental updates, I tried to bring something of the curious eye Latour & Woolgar brought while sociologists-in-residence at the Salk Institute in their (1979) *Laboratory Life*. Being able to witness on a regular basis the reciprocal interchange between the influence of existing ideas of memory traces in informing experiments, interpreting results and the usage of these interpreted results back to the creation of new scientific knowledge, was an immense privilege, but I do not profess that this grants me a more privileged place to comment on the science of memory. My attitude towards science follows Feyerabend (1975) in the sense that I understand science to be whatever scientists do, that there is no fixed scientific method scientists follow (least of all falsification), and I do not seek to fetishise or valorise the science I discuss as somehow a superior form of knowledge.

Secondly, after Hasok Chang, I believe that philosophy of science can do a form of "complementary science" (Chang, 2004), although I cannot say if I have satisfactorily achieved this here. I also broadly agree with Chemero's (2009) characterisation of cognitive science, to which I include the theoretical postulates made in neuroscience, to be immature or pre-paradigmatic and therefore requiring philosophy in tandem with experimental inquiry. Lastly, in terms of theoretical biology, my ideas have been shaped most strongly by my studies while an undergraduate biology student of the theories of Peter Mitchell, Lynn Margulis, and Stuart Kauffman.

As this thesis was being finalised, Mazviita Chirimutra's (2024) *The Brain Abstracted* was published. While I have not had the opportunity to engage with this book at the time of submission, I am aware that it too argues that

computational descriptions of brains break down because of their substrate-dependence. Any repurposing of the material in Chapter 3 (which deals with biological computation) for any future publication format would therefore engage with this work, in addition to the incorporating of feedback provided after the initial submission of this thesis.

Research Design

As a non-experimental thesis, there are different aims and methods. The ultimate goal was not to contribute a historical analysis of the concept of the engram or tabulate the various uses of the term in the literature or use case studies to track disagreements over its usage in science (both valuable projects). The ultimate goal was to *generate new ways to think about memory traces* in the most general possible sense. Given the lack of clear conceptual frameworks in the neuroscience of memory and how comparatively little treatment engram neuroscience has had, this meant having little precedent to start from. Thus, an *exploratory* methodology was chosen for this thesis. To maximise my coverage exploring potentially fruitful new ideas in looking to other disciplines, the project was divided into three separate streams.

Method A, which resulted in Chapter 1, was to reformulate existing relevant work in philosophy (primarily teleosemantics and contemporary philosophy of neuroscience) and theoretical biology so as to be relevant for the systems neuroscience of memory community. This also allowed me to better familiarise myself with existing theoretical work on representation, computation, and issued surrounding the concept of biological information to be used in Chapter 3.
Method B, which resulted in Chapter 3, was to do the reverse of Method A and bring cutting edge findings from the systems neuroscience of memory and articulate their relevance to a philosophy audience, in particular issues of representation. Doing so would force me to critically appraise the experimental studies I was interrogating in Chapter 1.

Method C, which resulted in Chapter 2, was to do something similar to Method B but bring the systems neuroscience of memory into dialogue with machine learning research and current trends in neuroscience-inspired AI.

This thesis is presented in the form of three separate pieces that can be read independently, rather than in the format of a continuous monograph (or baby-monograph, given this is not a PhD thesis) because each of these chapters is intended for different audiences (engram neuroscientists, neuroscience-curious machine-learning researchers, and philosophers of neuroscience, respectively). This tripartite methodology and structure was preferred over the writing of a single continuous work to maximise coverage of perspectives and cross fertilisation of ideas—in short to maximise coverage of different ways to think about memory and memory traces. I could not have started the project seeking to develop a unified conceptual framework or computational model of memory traces, since neither of these exist and it may be several decades before they do.

There was an intended in-built recursiveness to this to encourage cross-fertilisation. For example, although Chapter 3 is presented as the development of a view "with application to" the concept of the memory trace and representations, the view of biological computation I propose arose out of approximately two years of trying to conceptualise how memory traces, neural computation and representation could be understood together. In other words, it was not a separate project that has been applied to the subject matter of memory, it is the product of engaging with the concept of memory traces from a variety of different directions, including classical computers and connectionist models.

Overall, while this approach required covering a huge volume of present and historical material in drastically different fields, the overall exploratory nature of the research design meant I had the privilege to simply explore with provisional hypotheses, see what arose in the encounter, revise and repeat. This meant not just knowing in advance what connections and new ideas this approach would deliver, but surrendering any hope that it would deliver anything at all. I cannot be judge of whether it has succeeded in this aim.

Chapter 1 — If engrams are the answer, what is the question?

Abstract

Engram labelling and manipulation methodologies are now a staple of contemporary neuroscientific practice, giving the impression that the physical basis of engrams has been discovered. Despite enormous progress, engrams have not been clearly identified, and it is unclear what they should look like. There is an epistemic bias in engram neuroscience towards characterising biological changes, while neglecting the development of theory. However, the tools of engram biology are exciting precisely because they are not just an incremental step forward in understanding the mechanisms of plasticity and learning, but because they can be leveraged to inform theory on one of the fundamental mysteries in neuroscience—how and in what format the brain stores information. We do not propose such a theory here, as we first require an appreciation for what is lacking. We outline a selection of issues in four sections from theoretical biology and philosophy that engram biology and systems neuroscience generally should engage with in order to construct useful future theoretical frameworks. Specifically, what is it that engrams are supposed to explain? How do the different building blocks of the brain-wide engram come together? What exactly are these component parts? And what information do they carry, if they carry anything at all? Asking these questions is not purely the privilege of philosophy, but a key to informing scientific hypotheses that make the most of the experimental tools at our disposal. The risk for not engaging with these issues is high. Without a theory of what engrams are, what they do, and the wider computational processes they fit into, we may never know when they have been found.

Introduction

An outside observer might be forgiven for thinking that hidden in memory engram biology papers, or implicitly known among the systems neuroscience community, there is a clear understanding of what engrams are and how they relate to memory. Memory is defined at a behavioural level as changes resulting from past experience, and may also be defined as reconstructed internal representations (Dudai, 2007). In contrast, engram biology as a research program is dedicated to defining the biological basis of memory. Engrams are not memories, instead, they are what provide "the necessary (physical) conditions for a memory to emerge" (Moscovitch, 2007). Although not always, within neuroscience the terms memory trace and engram are used synonymously, with the term engram introduced in Semon's biological theory of memory (1904/21; 1909/1923). Inspired by Semon, in contemporary neuroscience engrams are operationalised as whatever physical or chemical changes meet four criteria: these changes are elicited by learning and persist, they can be reactivated ("ecphory"), that content is preserved from encoding to retrieval, and that between these processes the changes are not permanently active but dormant (Josselyn, Köhler & Franklin, 2015). Because cell ensembles can be experimentally labelled through co-opting activity-dependent immediate-early genes (IEGs) to express fluorescent and other proteins in those neurons activated during a particular learning experience, and because reactivating these specific labelled cell ensembles (but not others) often reproduces the original learned behaviour, and their silencing or ablation fails to reproduce that exact same behaviour, scientists have claimed these cells fulfil the function of the memory trace or engram and have appropriated and operationalised Semon's theory to help make sense of these experimental findings (Josselyn, Köhler & Franklin, 2015; Tonegawa *et al*., 2015a, Tonegawa *et al*., 2015b; Josselyn & Tonegawa 2020).

However, for all the pragmatic contemporary usage of the word, the engram remains an open-ended scientific construct. Semon's definition, restated in Tonegawa *et al.* (2015a) as the "*enduring physical and/or chemical changes that were elicited by learning and underlie the newly formed memory*

associations" is not tied to any specific biological level. It is telling that we must reach back over one hundred years for a conceptual framework to operationalise for current use in neuroscience[1]. In other words, although Semon's framework has been validated, it has not been elaborated, and there are no fleshed out alternative biological theories that would create healthy competition for the field. In general, the memory engram field remains under-theorised (Robins, 2023). Is this something that can be remedied by more advanced experimentation alone, or is *new theory* also needed?

For some, the concept of an engram may serve as a working abstract placeholder because the engrams that the scientific construct refers to have not been truly found, and the abstraction affords the elbow room required to manoeuvre while figuring out how engrams are realised in biological tissue. Despite experimental limitations regarding the temporal resolution of the cell-tagging protocol, differences in endogenous IEG function, controlling for off-target effects, and detecting exact differences between engram and non-engram cells, referring to the existence of "engram cells" is a pragmatic step forward based on progressive behavioural evidence that cannot be dismissed. Examining the cells involved in learning and recall did not *a priori* suggest that we would find cells labelled at encoding being reactivated at recall, or that these overlapping cells would be functional at the level of behaviour. Rather, this was demonstrated by empirical investigation (Reijmers *et al*., 2007, Liu *et al*., 2012; Garner *et al*., 2012). Meanwhile, the understanding of "how information is stored in an engram" is acknowledged as the major outstanding overarching question in the engram field (Queenan *et al*., 2017; Josselyn & Tonegawa, 2020) and cognitive science generally (Poeppel & Idsari, 2022).

For others, not only have engrams not been satisfactorily found, they may never be because the concept is not well defined (Sossin & Hardt, 2020) and there are views of memory that do not place prime importance on the existence of engrams (Schacter & Addis, 2007). If their existence can neither be proven or disproven, their scientific utility is questionable. Additionally, it is unclear from

[1] This is not to disregard the original theory's insightfulness, proper historical accreditation, or the very real sociological, ideological and experimental limitations that might have previously discouraged further theorising about the nature of the engram during the 20th century (Schacter, 2001).

either theory or empirical work what the information content is and how it is supposed to persist in the form or format of an engram. The language used in discussing existing engram theory can hide this lack of knowledge. To say that physical and/or chemical changes "underlie" memory is either a truism or insinuates that we have a clear theory for *what* the enduring changes are and *how* they do this underlying[2]. We have neither. Performing observational, gain-of-function, loss-of-function, and mimicry experiments has established *that* there is a replicable and specific linking between these unique cell ensembles and behaviour. However, when the theory being tested is ambiguous, such studies are limited in their power to explain what engrams are or *how* this specific linking works (Krakauer *et al*., 2017; Tonegawa *et al*., 2015b). There are now several putative levels for where the persistent changes that function as engrams may be found. These include i) synaptic weight plasticity (Martin, Grimwood & Morris, 2000; Kandel, 2001), ii) intracellular and molecular mechanisms such as protein or epigenetic modifications (Sacktor, 2011; Bédécarrats *et al*., 2018; Campbell & Wood, 2019; Gershman 2023), or iii) that the relevant mechanism is supra-cellular, such as changes in the structural connectivity of the connectome (Chklovskii, Mel & Svoboda, 2004; Ryan *et al*., 2021) or non-neuronal structures, like perineuronal nets (Tsien, 2013), astrocytes and oligodendrocytes (Kol & Goshen, 2020), or microglia (Wang *et al*., 2020). However, none of these proposals explain the *how*—how such changes are the appropriate mechanism capable of supporting recall of a given behaviour over any other[3]. This discrepancy—between the establishment of a linking between labelled neuronal ensembles (or any other level) and behaviour, but a dearth of understanding how this works—is a gauntlet that needs to be picked up by a new theoretical framework. However, our objective here is not to outline a concrete "engram theory". Developing new theories first requires understanding what our current framework is missing, including whole types of theory. This chapter therefore discusses current ambiguities in order to identify those areas any future theories must pay attention to.

---

[2] The use of such "filler-verbs" is a kind of interpretative (mal)practice not unique to engram biology but endemic to all disciplines of contemporary neuroscience (Krakauer *et al*., 2017).

[3] Our explanation may involve integration across many of these levels, but a multi-level explanation must still account for the *how*.

Section 1 addresses how there are in fact multiple things-to-be-explained when we are concerned with the biological basis of memory, and apparently competing explanations may not actually be in conflict. Section 2 singles out the dominant theory of engrams as operationalised in contemporary neuroscience, focusing on the limitations raised by its assumed modular architecture. Attention is paid to questioning the assumption that differences at the level of cell ensembles actually correspond to differences in the "type of information stored" by said ensembles. Section 3 argues this conflation occurs due to failure to dissociate questions about *information content* from questions about *vehicles of content*, and emphasises the need for more careful terminology regarding differing conceptions of what the engram should be. Section 4 is devoted to considering the limitations of theories of information in neuroscience and biology. For us, one of the most exciting aspects of engram technology is its potential to inform a new conception of information in the brain. Finally, in the Discussion section we summarise the implications of this approach for helping to guide experimental research.

Section 1 — Levels, Questions, & Explanations

1.1 — If engram is the answer, what was the question?

The major theoretical hurdle faced by engram biology is a lack of clarity regarding what the problems to be solved are, or not spelling them out in sufficient detail. The engram is assumed to be a good explanation, but what exactly are engrams posited to explain? The existence of something capable of functioning as an engram is usually invoked in order to explain the general phenomena of learning and memory. However, "learning and memory" is not a coherent, singular phenomenon, and when the phenomenon we want to explain is too general, multiple competing explanations will appear equally adequate. There are different questions that need to be parsed. In this section, three different

methods for doing so are presented based on i) biological organisation, ii) question-type, iii) and question-subdivision.

1.2 — Dividing explanations based on levels of biological organisation

Many biological processes and computations involve changes across multiple organismal levels. The changes studied in learning and memory research cover several orders of magnitude of various nested biological processes (Josselyn, Köhler, & Frankland, 2015; Craver, 2007, ch.5), from molecular to systems physiology. However, we are not sure which levels of organisation are most relevant for theorising about how the brain can be said to store information. It may be useful to draw analogies with similar problems in other areas of neuroscience. In their discussion comparing the relevance of neuron-and-circuit-level versus population-manifold-level descriptions when it comes to explaining cognitive phenomena, Barack & Krakauer (2021) illustrate the importance of choosing the best level of explanation with analogy to different answers to the question "why did the window break?" The *first-level explainer* (we denote here as E_1) should be "because the ball was thrown at it" and not explanations based on the physiology of arm-throwing or the Si-O amorphous molecular structure of glass. These *second and third level explainers* (E_2 and E_3) are defined by their ability to explain E_1, but that does not make E_2 and E_3 better answers to the original question[4].

If we are seeking to explain the biological basis of memory, we are not seeking to explain "the engram". Engrams may be posited as a first-level explanation (E_1) *for* the biological basis of memory, or may be a given step (E_n) along the way, say E_5. Alternatively, the engram may refer to a given chain of processes, with no one privileged level. Since we do not know what the engram is, our focus shifts to picking out the next level of explanation down (E_{n-1}) to that which might support the engram. Thus, when a biological mechanism or process is conjectured as an explanation, care should be taken to ascertain whether this is

[4] If the question is not fixed but we vary it, then what is E_2 to one question may become the E_1 for another.

an explanation for the biological basis of memory, or actually the-thing-that-explains-the-thing (etc.) that explains the biological basis of memory[5].

Although this seems like a straightforward pitfall to avoid, it is not when it comes to memory. This is because the problem to which memory is a solution to, that of dealing with dynamically changing external conditions that have some statistical dependencies, is not a problem that manifests once at a single scale[6]. It is evolutionarily universal, and solutions or adaptations have manifested across multiple levels of biological organisation and phylogeny (unlike, say, speech production or motor imagery). Bacteria have been said to store information in protein post-translational modifications (Sterling & Laughlin, 2015; Gershman, 2023) and slime moulds are capable of anticipating temperature changes (Saigusa *et al.,* 2008). Should any biological mechanism of "storing information" by modifying some structure be considered an engram? Avoiding a situation of "endless engrams" all the way down, in which vastly different processes all stake claim to the same name, implying some non-trivial equivalence across all, is preferable to us[7,8]. Granted, different biological processes across the evolutionary continuum and in parallel across multiple scales within multicellular organisms involve the creation of specific changes could be said to function as engrams that are relevant for various kinds of phenotypic plasticity[9]. However, the question to be addressed in order to explain behavioural memory is *which aspect of the problem is each mechanism or process a solution to* (see Figure 1) and how these different information storage mechanisms are hierarchically scaffolded together

---

[5] The pitfall to avoid is to take, say, E_{20} for E_2.

[6] This question-answer dialectic of an environment posing problems to which an organism evolves "solutions" is imperfect (see Lewontin, 1983) but illustrative for our purposes here.

[7] There may be nothing that unites them other than uncertainty reduction very generally defined.

[8] Biological processes are nested and previous solutions are repurposed, but this does not make a solution at one level universal or that it will be used for addressing the same problem in future evolution (Gould & Vrba 1982; Kauffman, 2019). Solutions to general evolutionary problems can be convergent in some cases, but they can also be divergent. When the same problem is faced at different levels of organisation, even within the organism, divergent solutions are found. Both we and our immune cells have to "navigate." We use a musculoskeletal system and the algorithm of walking. Immune cells use cytoskeletal rearrangements and roll along tissue with surface proteins. Like the general problem of navigation, it is sensible to assume the similarly general problem of learning and the solution of storing changes also has many levels.

[9] For example, DNA, immune memory, and external memory (Donald, 1991; Clark & Chalmers, 1998)

(Simon, 2002; Hoffmeyer, 2007). Experimental data without theory does not reveal this[10].

Given these general considerations, competing proposals for the location of "the engram" suddenly no longer seem to be in competition at all. In recognising that synapses (or synaptic strength at least) may not be the site of the lasting changes that count as an engram (Chen *et al.*, 2014; Ryan *et al.,* 2015), two theories, one that goes down a level of biological organisation from synapses to nucleotides and protein conformational states (Gallistel, 2017; Gershman, 2023), and one that goes up a level to structural connectivity (Chklovskii, Mel & Svoboda, 2004; Tonegawa *et al*., 2015b; Ryan *et al*., 2021), have seen a resurgence or been proposed (for a detailed assessment, see Colaço & Najenson, 2023). However, the spatial scales of causal influence implicated at each level may differ dramatically. It may be completely accurate to claim that there is information *in* neurons, that it is molecular and that stable distributed molecular states could function as engrams, but these may be engrams *for* the cell informing itself about, say, cytoskeletal rearrangements, and not engrams directly *for* whole-organism behaviour that is retrievable over millisecond timescales[11]. The former is unlikely to be an E_1 for the biological basis of how an organism such as a mammal is capable of creating an episodic memory. It may feature in a future theory of the biological basis of episodic memory at E_{10}, say, and may indeed be necessary, but E_{10} alone is insufficient.

In contrast, it is also possible that the engram will not be the E_1 for the biological basis of episodic memory but will require it being situated in a larger context or a chain of explanations. A "manifold account of the engram," for example, might include explanations linking the disinhibition of a previously latent connectivity pattern with neural population activity taking on a different dynamical trajectory in a certain context (Langdon, Genkin & Engel, 2023). The connectivity pattern may in turn be explained by cytoskeletal rearrangements and other molecular level changes, just as motor neuron firing can at a certain level be

[10] Although we might claim that the level of learning and memory we study in systems neuroscience is tethered to whole-organism behavioural parameters, and all experimental data is collapsed relative to this, the changes we choose to look at, stain, sequence, or quantify are not determined by the data but by hypotheses of where to look.

[11] Stable molecular states may be E_{n+2} explanations for cytoskeletal rearrangements, E_{n+1}, which are explanations for E_n, but this does not make E_{n+2} an explanation for E_n.

said to implement throwing a ball at a window. Spelling the problem(s) out at different gradations of complexity will facilitate selecting the appropriate type of engram or information storage mechanism posited in a given hierarchy.

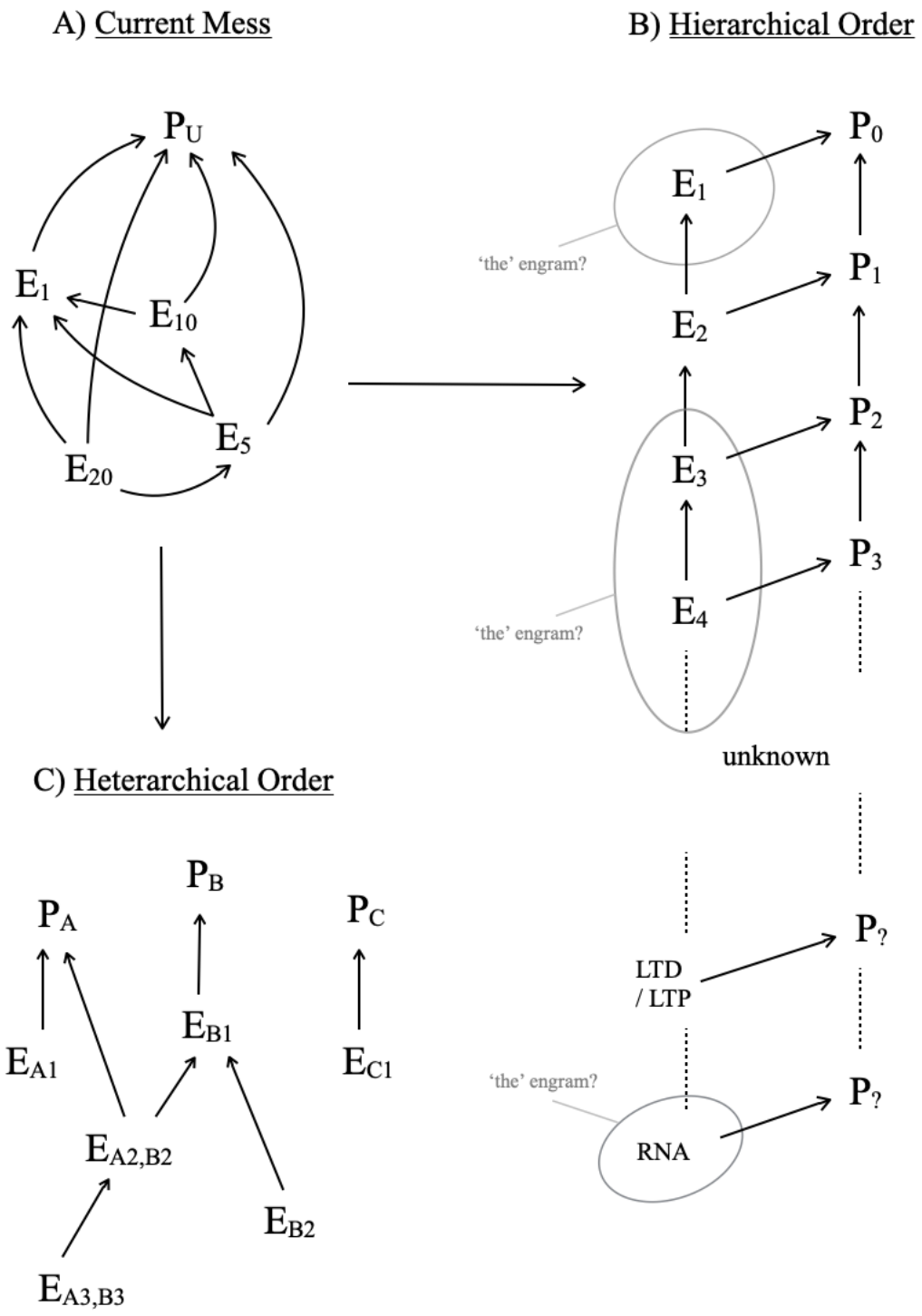


Figure 1: A) When the problem to which engrams are posited to solve is articulated over-generally as a kind of universal problem, P_U, multiple biological mechanisms may seem to qualify as equally adequate explanations. Either many biological mechanisms get to be engrams or their conglomeration is an engram, neither of which is illuminative. In moving from A) to B) or C) by unfolding P_U into a series of more specific problems that have to be solved, the biological mechanisms can be parsed and scaffolded into a more coherent theory, whether hierarchical or heterarchical. We may choose to define the engram as a particular causal motif (Ross, 2021) or as that mechanism which offers the greatest control in terms of causal manipulation of memory content, or as a chain of biological mechanisms, none of which alone qualify as an engram. Specifying problems discretely rarely works in practice, and here only illustrates a conceptual point.
LTP = Long Term Potentiation, LTD = Long Term Depression.

1.3 — Dividing explanations based on question-type

So far we have discussed choosing explanations with respect to a fixed question type (and by varying the level at which the question manifests), but we can also vary the type of question we ask. For example, classic heuristics like Tinbergen's (1963) "Four Questions" for any given behaviour differentiates *how-questions* (the development of a behaviour within an individual lifespan, and its mechanism of function) from *why-questions* (the evolutionary history of the behaviour and what adaptation it affords). We used a why-question in asking why the window broke, but if we ask "how did the window break?" then explanations for the brittleness of glass and velocity of the ball are indeed the relevant form of explanation. It could be argued that when it comes to the memory relevant for behaviour, the *what* (the behaviour we are studying) is learning about the social and physical environment, the *why* is recall in the future for adaptive behaviour, and the *how* is the engram and its expression. However, there is no singular answer to the how-question, because *how to form* an engram versus *how to use* an engram we presume are different processes. Formation and reactivation may

therefore require different levels of description. For example, an explanation for *how an engram is used* may require situating structural changes in terms of their downstream effects (e.g. funnelling network activity into a subspace regardless of where the network starts at the time of retrieval, but which preserves some high-level invariant properties with the landscape at the time of encoding), before jumping to explaining behaviour.

1.4 — Dividing explanations based on subdividing the problem

Section 1.2 introduced the problem of potentially many different processes being collapsed into an unknown hierarchical relevance with respect to a particular observed behaviour. Section 1.3 demonstrated different kinds of questions. Drawing on both concerns, this section focuses on taking the general problem of "dealing with dynamically changing external conditions…" and decomposing it into sub-problems. For example, we might focus on the problem of efficient information storage (Sterling & Laughlin, 2015), or the problem of making stable changes accessible (Ryan & Frankland, 2022), or the problem of whether a given type of engram is computationally content-addressable or address-addressable (Gallistel, 2017), to speculate on a few. The most relevant level description may differ for each, and there may be multiple versions of each problem across scales. Deciding how to divide a problem like biological information storage into subproblems is also not something that can be approached solely as an engineering problem, but requires consideration of the phylogenetic history of the organism and its ancestral environments (Cisek, 2019).

Failure to decompose the problem hinders inferring which changes in what process are most relevant. If all we are looking for is structured variation in a particular behavioural parameter, there is an endless array of variables across the scales of biological organisation that we can intervene with that will produce such effects. This risk, as raised in Jonas & Kording's (2017) problem of inferring the logic of a microprocessor using standard neuroscientific gain-of-function and loss-of-function causal interventional techniques, is that we will take the wrong level of explanation to be E_1 just because manipulation at that level resulted in

structured outcomes in behavioural parameters. Structured variation in behavioural outcomes due to protein synthesis inhibition (e.g. no memory consolidation, see Ryan *et al.*, 2015) does not necessarily mean that our E_1 for how an engram instantiates a particular form of information will involve describing a given protein's function.

Section 2 —What flavour of model are we using in engram biology?

2.1 — Current engram theory is modular and compositional

The existence of multiple levels of engrams aside, the dominant sense of the engram that has emerged in contemporary systems neuroscience research is that the engram is somehow instantiated as cellular ensembles and a brain-wide distributed complex (Josselyn, Frankland & Köhler, 2015; Tonegawa *et al*., 2015; Josselyn & Tonegawa, 2020; Wheeler *et al*., 2013; Vetere *et al*., 2017; Roy *et al*., 2022). But exactly what kind of model is being proposed? Must the engram necessarily be a global and non-decomposable unitary entity, or can discrete "sub-engrams" have fixed identities while still making up a "super-engram" or "engrome," and is this the most appropriate way to carve up the system? Having a clearer understanding of what assumptions this model entails will aid in refining or replacing the existing paradigm.

The terms engram cells, engram cell pathways, engram components, and engram complexes were defined and proposed in Tonegawa *et al.* (2015a) in a contemporary attempt to operationalise Semon's engram theory in light of the last century of advances in biology. In short, there are ensembles of cells (engram cells) that instantiate different components of a much larger, distributed engram (the engram complex) joined together by connectivity pathways between these engram cell subpopulations (Josselyn, Frankland & Köhler, 2015; Josselyn & Tonegawa, 2020). Each of these different sub-engrams or engram building-blocks contribute different pieces of information, such as hippocampal DG and CA3 for contextual information and CA1 for contextual and temporal information

(MacDonald *et al.*, 2011; Tonegawa *et al*., 2015; Roy *et al.*, 2022). A further level of complexity is added by the existence of inhibitory neuronal ensembles ("inhibitory engrams") contributing to broader engram complexes which have been proposed as a general feature of brain function (Baron *et al.*, 2017; Koolschijn *et al*., 2019; Sun *et al*., 2020; Nambu *et al.*, 2022). The picture that emerges is one with a strong sense of modularity and compositionality, where parts of specialised function can be viewed as plug-and-playable modules, participating in a larger distributed engram which in sum we presume has sufficient capacity and richness of stored information to reconstitute a memory.

If explicitly articulated, a modular or compositional theory may serve as a flexible and helpful architecture to explore. However, there are different ways of treating the parts of any modular theory. An *essentialist* approach explains the behaviour of a system by appealing to the intrinsic power and casual abilities that inhere within its parts. These elements have necessary properties that make them what they are. In contrast, a *relational* approach explains the behaviour of a system by appealing to the unique way the parts are configured with respect to each other (Buzsaki, 2006). Another way of describing these positions is as follows: *what confers on the trace the properties that it has, the intrinsic qualities of trace, or the system in which it is embedded?* We explore the differing consequences of essentialist and relational approaches to the basic modular theory of engram biology across different levels of granularity[12].

2.2 — Relational and essentialist conceptions of 'the engram'

We start with the foundational construct of the engram first before talking of cell-ensembles and ensemble-complexes. An essentialist approach views the engram *as* or distributed *in* the constellation of cells that were activated during learning and whose reactivation produces a behaviour indicative of recall. Any cells that were active during learning that did not get incorporated into the ensemble, or

[12] We do not have to limit ourselves to two positions. See Lalpane (2016) for a fourfold classification of "stemness" in stem cell research. Additionally, we also do not focus on hypothesis that the engram is molecular in this section in the interest of space. However, there is nothing that precludes imagining differences in relational or essentialist approaches at a molecular level.

neurons not active during learning but were recruited later, are an unexplained nuisance. Information is instantiated in a fixed, solid entity, and any alterations will change or corrupt the information preserved. So-called "representational drift" (Rule, O'Leary & Harvey, 2019) might therefore be taken to be a threat to the integrity of the stored information.

A relational approach to defining the engram does not deny that these neurons must play an important part, but attempts to appreciate their role in a wider context. We summarise this attitude with the maxim *it's not what the changes are, it's what the changes do.* For example, Guskjolen and Cembrowski (2023) recommend that the construct of the engram must expand to include both non-neuronal cells and also those neurons that are actively silenced during learning and must be re-silenced for retrieval. In other words, knowing which cells are inactive is just as important as those that are active for information storage and retrieval, because plasticity in engram cells and their connectivity patterns will change the local connectome around them. Unlike artificial networks with scalar activations, excitatory/inhibitory balance is an essential principle of nervous system function, where computationally inhibition can be understood as transiently modifying the morphology of other neurons (Buzsaki, 2006). Whether we consider a necessary unity between excitatory and inhibitory cells to be required for something to count as an engram, or we treat excitatory and inhibitory ensembles to be dissociable structures that cooperate, it is clear that any theory of engrams must account for the relevant context that makes an engram possible. Just as a gene is understood not merely as the sequence of base pairs that code for a protein but a region that includes enhancers, promotors, introns, the presence and absence of general and specific transcription factors, so too is there ample room to speculate about the neurobiological context that makes something an engram, or makes it the engram that it is. This sets us up to consider engrams not as monolithic entities or "data structures" of one format, such as a singular "information molecule" or singular kind of information (e.g. binary states) realised in several different biomolecules or larger structures, but rather complex (that is, non-decomposable) informational units that may have many necessary pieces, none of which on their own are sufficient to count as a piece of latent information. But what exactly are these "pieces of information"? Do differences at

a cellular, ensemble, or complex level explain the differences in the information content held by one engram versus another?

2.3 — Ensembles are conflated with pieces of information

A common assumption, introduced in Section 2.1, is that different cell-ensembles each instantiate different pieces of information to a brain-wide engram. This encompasses any statements to the effect that one ensemble is said to provide spatial context, another temporal, perhaps a valence component from the amygdala and the various sensory modality engram components from sensory areas, and so forth. However, it is a strong assumption that differences at the ensemble-level are the *right level of explanation* for differences in the "type of information content" the hypothetical engram is posited as preserving. Relating these is a potential category error (Ryle, 1949) and cannot be assumed. It is not yet clear that cell populations in different areas with different functions somehow neatly instantiate the informational components that make up the full engram. It is a promising hypothesis, but must be articulated as a hypothesis and not an assumption, as it still lacks experimental evidence. There are at least two items this hypothesis must be able to account for, i) *Intra-ensemble differences:* What is it *about* one ensemble of a given type that differentiates it from another of the same type? For example, temporal engram component X versus temporal engram component Y in CA1. And ii) *Inter-ensemble differences:* What is it *about* one type of ensemble that makes it different from another type? A dentate gyrus engram ensemble differs from a medial prefrontal cortex engram ensemble, but in what way are they the same and in what ways do they differ? Is it region alone? Ideally, we want a taxonomy of engram types that avoids the mistake of grouping heterogeneous things under a unifying construct (lumping error), while also avoiding unnecessary fragmentation of engram-ensemble types into as many as there are possible ensembles (splitting error). The essentialist-relational distinction can be used to draw out questions from these ambiguities.

2.4 — Essentialist and relational views of ensemble and complex function

At an *inter-ensemble* level, there are three essentialist ways of explaining what makes one type of ensemble instantiate the information that it does. First, perhaps it is the *type of changes* undergone (whether intracellular or membrane localised) that confers the informational specificity of that engrams function (1). Given two approximately identical ensembles in different locations, the unique changes undergone determine, for example, whether the engram is a fear engram, because the cells underwent "fear changes." The second is that all engram cells undergo the *same kinds of changes*, but it is the cells in which the changes occur that confers informational specificity (2). Since the region does not itself confer properties, either it is something about the intrinsic properties of the cells in that region (2.i), the connectivity of the cells in that region (2.ii), or a combination of both (2.iii). The third possibility meets (1) and (2) somewhere in the middle: there are different kinds of changes, but location matters too (3). For example, Shpokayte *et al.* (2022) demonstrated topographic segregation (with some overlap) between ventral hippocampal engram cells for aversive and appetitive experiences, as well as different transcriptomic profiles for each population. Whatever the specific changes are in each case, they denote either the positive or negative aspect of the experience. Regardless of which account of how the type of information is determined at the level of the engram ensemble, an essentialist explanation for how these ensembles work together is straightforward—their place in the brain-wide engram-complex is merely additive[13]. Failure to recruit one ensemble still results in a functional memory, minus, say, knowledge of when an experience happened.

A relational approach would argue that, helpful as the three distinctions are, these changes are not explanations for why one ensemble instantiates a given type of latent information over any other. The changes as described do not explain *how* it is that this ensemble can store a "spatial type of information" while a given other ensemble is capable of "positive valence information storage." We want to know what confers this stored informational capacity and type, and if changes at

[13] Ensemble-type$_1$ + Ensemble-type$_2$ + Ensemble-type$_3$ = successful retrieval of memory to guide behaviour.

the level of the ensemble even are the most relevant level for determining the type of latent information. A relational approach would also be open to the possibility that information is not determined by persistent changes at the level of cell ensembles, but only at the level of the brain-wide engram when activated. In contrast to an essentialist view of engram components as fully formed cogs that are simply "called up" by context with appropriate syntactic synchrony, a relational approach might argue that each component is mutually determined, or has a strong dependence on the other components present based on how they are arranged within the connectome (Vetere *et al.*, 2017)[14]. This view informs how we test questions about whether or not the cell ensemble is the best level to locate discrete storage arising, whether there is a minimal number of ensembles needed[15], and whether or not ensemble parsing and their syntactic activation order (Buzski, 2010) must be taken into consideration for determining informational specificity at a brain-wide level for successful recall.

2.5 — Cutting across the levels

To conclude Section 2, the struggle to flesh out the details of the modular framework go hand in hand with a lack of clear hypotheses regarding the basic concept of the engram itself. It is likely that the way we have divided the problems faced at each level is misplaced. Some problems may apply across all levels, and to solve the problem at one level requires solving them across all.

---

[14] In linguistics and philosophy, this is called semantic inferentialism or conceptual role semantics. In fusional but especially polysynthetic languages, morphemes (the units of meaning) when split apart often do not make sense on their own, thus communicating involves using sentence-words. A missing morpheme is easily guessed (pattern completion).

[15] Say a protein may be functional only with 8 subunits of different isoforms. If only 7 subunits are transcribed, the protein is not coherent, either it does not form, or forms something else with a different function.

Section 3 — If engrams are vehicles, what are they vehicles of?

3.1 — Representations vs neural representations

In this section we introduce a common if controversial distinction made in philosophy between the *vehicle* and *content* of mental representations (Dennett, 1978; Hurley, 1998). Content or semantic content is what a representation is about, such as a particular scene or scent in the case of an immediate perceptual representation and/or an indirect uncoupled conceptual representation such as an idea, abstract concept, or mental map that we use to stand in and flexibly guide decisions in the absence of direct experience. The vehicle is often taken to be a physical (e.g. neural) part or process (Shea, 2018; Burnston, 2021). There is much debate over what conditions must be satisfied for something to count as a representation (Ramsey, 2007; Krakauer, 2022), and whether they map neatly to discrete vehicles or not. However, this debate on representations is typically agnostic as to neural implementation. Thus, both senses of representation just outlined (immediate and uncoupled) are very different senses of representation than when neuroscientists refer to an activity pattern as a "neural representation" of a face or tuning curves or firing rates as *representing* some features of a stimulus (e.g. Chang & Tsao, 2017). This is a softer concept (that of encoding) and one we will assess in section 4. For now we need only note that neural representations are unlikely to qualify as representations in the sense outlined above. Instead, they may be candidate vehicles of representations, or recording-apparatus-relative proxies of vehicles at best[16]. Nonetheless, employing the vehicle/content distinction to engrams is instructive, as it can help us separate out at least three categories of engram concepts with their own sub-types. These three are *engram-as-content, engram-as-vehicle,* and *engram-as-ensemble*.

[16] There are more pragmatic accounts of what it means for something to be a representation in practice, under which the following discussion may qualify as treating engrams as representations (Cao, 2022). We do not adopt this pragmatism here for the purposes of avoiding conflation between memories and engrams.

3.2 — Engrams-as-content

As similarly outlined in Gershman (2023) and Najenson (2021), understanding the biological basis of memory encompasses both understanding memory content and how it is instantiated by some physical vehicle. The first conception is that the engram refers only to the content. Although this is not the most common conception of the engram, it is worth addressing. Within this, there are two senses of content that need to be distinguished—representational and non-representational content. Conceiving of engrams as representations includes viewing them as downsampled and/or "encoded" versions of memories (Moscovitch, 2007).[17] However, memories (at ecphory) may also be referred to as representations (Dudai, 2007). This is a view of memory representations as "the same thing" transforming from latent trace to active and back again, flip-flopping between two states. While we colloquially refer to the contents of an image and the contents of a file using the same word *contents*, engrams were never envisioned to be *of the same stuff* as memory, hence their separation from ecphory (Robins, 2018). The second sense of engrams-as-content is that engrams are not representations with rich semantic content comparable to human concepts but some form of simple informational content locally for the neural system itself. For example, this includes the idea that an engram (or a hippocampal engram) is a particular *index* for constructing a given memory (Goode *et al*., 2020), meaning the informational content of these engrams are addresses. For example, this includes the idea that an engram *is* a particular index for constructing a given memory. Talking of information rather than representation suggests that there is a difference in format between engram and memory so large it may require a whole other way of conceiving of engrams other than as the "encoded version" of a memory. This is an idea we will return to, but requires clarification.

[17] This differs from the sense of representations outlined in the previous paragraph, which pertains to first-person deliberative use of representations (Krakauer, 2022).

3.3 — Engrams-as-vehicles

As per Semon and recent definitions, the engram is envisioned as an *implementational* construct that concerns physical and chemical biological changes. However, we often fail to distinguish between the conception of *engram-as-content* and *engram-as-vehicle*. There are at least three possible senses of *engram-as-vehicle*, based on what the engram is a vehicle of or *for*. First is that engrams are not memories but the neural vehicles of memories. However, whatever the neural implementation of memory is, this concerns what happens at time of recall. The engram is a separate neural implementation which persists even when a memory is not being created. Engrams, therefore, do not have to be the vehicles of memories[18]. Given this, we may propose a second though similar option—that engrams are not the vehicles of memory but are vehicles instantiating a stored, reformatted, and downsampled encoding of a memory. This does not abet the problems outlined with the *engram-as-representation* view above. Finally, there is the conception that engrams are not vehicles of representations but vehicles of a simpler informational content. The content may not be seen as semantic in the representational sense of being about or standing in for things outside the brain, rather the content of the vehicle is what it does: for an engram theory that places primacy on connectivity, it may be shifting the probability of a given neural activity pattern being expressed through altering the pathways that neural activity is capable of taking. At this point in time, it may no longer be fruitful to look at the engram from separate *vehicle* and *content* perspectives, or talk of content at all (Figure 2). This is the view we are sympathetic to and wish to see clarified further in the context of engram biology.

3.4 — Engrams-as-ensembles

The last category of engram views is that engrams have already been discovered and they are unique sparse ensembles that we can partially label and manipulate.

[18] This conflation may come from using "memory" to refer ambiguously to both memory and memory traces.

These cell-ensembles may or may not qualify as vehicles and could overlap the previous section (see Figure 2). However, if these ensembles are not the vehicles of stored information, then these engrams do not fulfil the objective of explaining the biological basis of memory and we will need to posit some other construct that fulfils this role. If these ensembles are indeed the vehicles, then we need to be able to bring to the foreground those most relevant changes and relegate those not relevant but still picked up by the tagging protocol. However, current experimental methods have not revealed to us what these most relevant changes are (are they molecular changes, connectivity changes, or something else?). Although adopting the engrams-are-ensembles definition may seem pragmatic, it gives the illusion that the research program of engram biology has finished answering its fundamental question and moved on to simply looking for applications to behavioural neuroscience. Although the existing techniques of engram biology are powerful and will undoubtedly lead to more applications where a true understanding of the biological basis of memory is not essential (for example, in understanding the biology of forgetting, addiction, trauma, delusions, etc.), if we want the engram to explain the specificity of memory, then we must admit that satisfying Semon's criteria is only a beginning and is not alone sufficient (Ortega-de San Luis & Ryan, 2022). In other words, we have not yet isolated the engram. Rather, these labelled ensembles may encapsulate, be adjacent to, or partially instantiate the "enduring physical/chemical changes" that constitute the engram[19].

[19] Different IEG expression profiles do not always overlap, leading to the confusion that two disjunct labelled populations may both be called engrams. At best, differences in IEG function may be for implementing preservation of different types or layers of information that ultimately comprise the engram (Nambu *et al*., 2022).

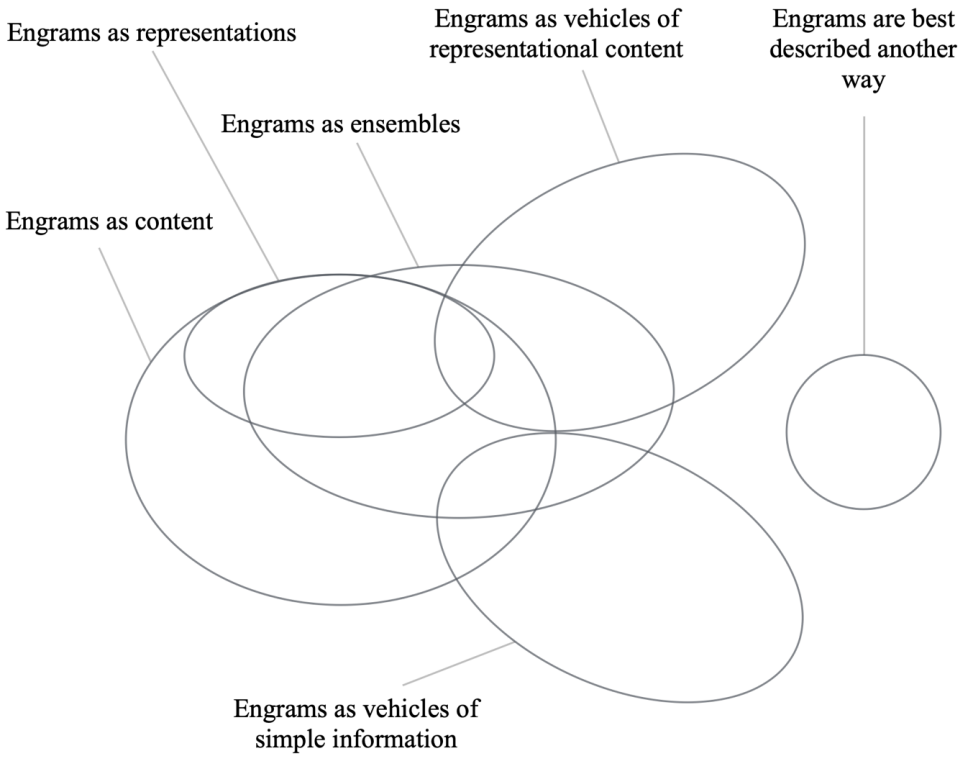


Figure 2: Different overlapping conceptions of what engrams should be.

3.5 — A vehicle-information distinction

These distinctions are controversial[20] and coarsely wielded, but entertaining these differences opens a space for considering other possibilities. *If engrams do not implement storage of representations, what do they have, or what can they be said to function as?* Applying the vehicle-content distinction dissociates different senses of the word engram, but this is not its only utility. Even if engrams do not have representational content, there may be utility in employing a modified *vehicle-information* distinction, whereby "information…persists via a memory trace, which is the vehicle for information preservation" (Mace, 2021). There are two avenues leveraging this distinction helps us consider. Just as we presume that not anything can count as a representational vehicle, we presume that not everything is *competent* to function as an engram. What would it take for a vehicle

[20] To distinguish between latent information and vehicles is not to separate them and treat them as separate entities, but to look at the same phenomena from different perspectives. Nonetheless, it creates a dualistic divide that is not always desirable (Mitchell, 2023).

to function as preserving information and what criteria would exclude a biological structure from qualifying as competent (e.g. scar tissue preserves persistent changes despite cellular turnover)? This means further refining the four criteria for what it takes for something to count as an engram (see Josselyn, Köhler & Frankland, 2015). Second, this distinction makes clear an asymmetrical bias in engram biology research towards characterising the vehicle of stored information. Discoveries about the vehicle, even figuring out the mechanism that makes something an engram, does not necessarily tell us anything about what information is stored. As will be discussed in Section 4, we have a poor understanding of what this latent information may be. At the very least, we can use the vehicle-information distinction to check ourselves, asking "do the questions addressed in this study focus solely on the vehicle, or do they also tell us exactly what information is stored?"

This has precedent in biology, as a clear vehicle-information distinction is seen in genetics. A gene may be given a physicochemical description as a string of nucleic acids with certain molecular properties, or an informational description as a *symbolic* sequence that has a particular meaning in the context of the cell. It can be interpreted though specific cellular processes into a functional peptide or nucleic acid. DNA/genes lead a "double life" as matter and *semiotic symbol* (Gazzaniga, 2018; Pattee 1987/2012). In the case of genetics, an entirely new way of thinking about biology had to be created, that of thinking in terms of a symbolic code system: The eventual cracking of the DNA triplet codon to tRNA amino acid allocation, proved that the mapping is arbitrary and therefore code-like, with redundancy. This shift required a level of abstraction that turned genetics from a form of molecular biology into a kind of biological information science.

Important for our discussion is that information does not fall out of physical knowledge of the vehicle. The physical form of two genes with the same base-pair length may be near-identical, but what the genes code for may differ radically (a plant enzyme or a SARS-CoV-2 cell surface protein). The meaning can be divergent even from the same information in cases of differential splicing and post-translational modifications. Whether we are looking at DNA with the fine resolution of an electron microscope or a chromosomal karyotype, we are not

able to see this genetic information. We are characterising its vehicle. When we look at a labelled cell ensemble in a brain tissue slice, we are looking only at one side of the engram. Distinguishing between engram and non- or other-engram cells is not looking at stored information. Double-blinded to a tissue slice of labelled cells, we could not say "this is a fear engram" or "this is an episodic memory from this place and time." Again, *knowledge of the vehicle of information is not enough to tell us what this engram is for* (Eichenbaum, 2016). In the case of engram neuroscience, we may even have the resolution needed, but cannot from the forest pick out i) what the vehicle is, or ii) how it affords the capacity for the specificity of memory and therefore behaviour. These are two different targets of explanation; analogously, explaining the structure of DNA is not the same as explaining the role of DNA in heredity and evolution. Regarding i), identifying the vehicle is not a binary YES/NO achievement, since there are levels of precision. When Avery, MacLeod & McCarty (1944) and Hershy & Chase (1952) discovered DNA to be the molecule responsible for transmitting genetic information and not protein, they had identified the vehicle in a coarse sense. This did not tell us what the genetic information was, or reveal which changes in the physical vehicle were most relevant for storing information. When we find that unique labelled cell-ensembles are capable of reinstating a behaviour, we may (at best) have identified engrams in a similarly coarse sense. Additionally, we would not say that a gene stores representations of proteins, organelles or organs. Rather there is a stable molecule (DNA) that constrains the creation of functional molecules in a highly reproducible way using an arbitrary code-like mapping.. Similarly, we are not suggesting we try to "find the information in the brain" in a literal sense, rather we are arguing that engram biology needs a framework that makes a similarly abstract leap to help us understand whatever constructs serve as the biological basis of memory. We feel a vehicle-information distinction is important, because it illustrates how engram biology is focused exclusively on the vehicle side of the engram and how there is a dark side of the moon to engram biology that is rarely discussed. With this in mind, we now focus on how we might understand this information.

Section 4 — Information and coding

4.1 — The necessity of a theory of information

Talk of information-processing is ubiquitous throughout neuroscience, but what is meant by information is rarely specified. This same state of affairs applies in engram biology, where understanding the biological basis of memory is often framed in terms of discovering how information is stored in the brain and manages to persist through time (Josselyn & Tonegawa, 2020). But what exactly is this information? The use of information-concepts in biology is controversial (Godfrey-Smith & Sterelny, 2016). However, in engram biology there are behavioural and other phenomena that may be difficult to explain without invoking the existence of stored information. For example, various forms of memory learned during the infantile amnesia window are not lost but rather inaccessible and can be artificially retrieved (Guskjolen et al., 2018; Power et al., 2023), and other apparently lost memories, including in Alzheimer's disease models, can also be recalled (Ryan *et al*., 2015; Roy *et al.*, 2016; Perusini *et al*., 2017; Abdou et al., 2018; Poll *et al.*, 2020; Lei *et al*., 2022; Bolsius *et al*., 2023; Autore *et al*., 2023; O'Leary *et al*., 2023). Accounting for these phenomena without positing latent information, or another concept that performs a related function, is problematic. To the extent that it is helpful to posit the existence of information and use information-concepts, a stronger theory of what this information is is needed (Robins, 2023). We therefore survey two different existing general conceptions of information, and conclude that neither are completely satisfactory for use in engram biology. Instead, the techniques of engram biology mandate creating a new conception of information for use in neuroscience.

4.2 — Classical information theory and coding metaphors

The first general class of views on information are inflected by classical information theory and computer science (Stone, 2022). Information is treated as something entering the brain from an external source (e.g. a stimulus) whereupon it is "encoded" (Roediger, 2007), and is passed around between neurons, perhaps travelling in inter-spike intervals as a rate or combinatorial code (Gallistel, 2017). Learning here is the acquisition and consolidation of relevant information extracted from "sense data," with stored information held in something like a weight matrix (Luo, 2020). However, classical information theory is not a theory of information content but of quantifying how information can be optimally transmitted. It explicitly excludes accounting for the semantics, or meaning, of what is communicated (whether termed a message, signal, or information)[21]. Neuroscientific theories about how information is coded, the computational and therefore metabolic benefits of using coding principles (Friston, 2010), inherit this same problem. Although the details of such views can differ, as not everyone might agree that the brain computes using *bits* (Sterling & Laughlin, 2015), or uses a read-write computing architecture (Gallistel & King, 2010), they do not provide an account for what the information being coded is. This leaves much to be desired, as neuroscientists often care about the meaning of what brain signals are communicating.

The limitations of this heavily abused coding metaphor are summarised in Brette (2019). Principally, when we look for a correlation between any external variable, whether sensory or higher order task structure, and neural activity (e.g. Ca^{2+} events or Blood Oxygenation Level Dependent patterns as proximate measurements), we have assumed the existence of information, but only in our capacity as a third person observer fixing a relation between two variables, one inside and one outside the brain. The brain does not have this third-person perspective, and so this cannot be assumed to be the information stored in engrams. Instead, for any future engram theory seeking inspiration from computer science, it may be more fruitful to direct attention to computer hardware engineering principles.

[21] In telecommunications and computer science, this is not a problem, as it is humans who design the hardware and determine the arbitrary conventions for how variables should be encoded for other humans to decode them.

4.3 — Information is relational and momentary

A second view, one that tries to accommodate the meaning of information, is that information is not a noun, property or thing that can ever be localised, but is inherently relational or contextual. There is nothing *in* DNA. Nucleic acids like DNA and RNA do not *carry* information by themselves. Outside the context of the cell, they are merely polynucleotides. There is only information *for* a given system in context. Neurotensin, interleukin-17, insulin, and serotonin only have biological meaning or unique consequences in the context of signalling, and are differentially responded to by cells on the basis of the receptors they express. The meaning of a signal has physical constraints, such as molecular mass and hydrophilicity, but these constraints do not necessarily determine the meaning, which is fixed by evolutionary convention and may be modified by learning. Information in biology is not something that inheres to any physical pattern, it is information only in relation to or for some system in context (Brette, 2019; Buzsaki, 2006; Deacon, 2021). This means that information can never be something held or stored, but is being continually created. In the context of theories of memory, this may fit with a view of information only existing at the time of recall (Schacter & Addis, 2007). Under this view, the function of the engram (if posited) is not to store information, but to participate in the process of reproducing information meaningful for action in the relevant context. There is none of this information *in* the unique pattern of persistent changes that make an engram. Although viewing recall as a process of forming transiently existing information *using* engrams fits with Semon's account of ecphory, this view still leaves the biological basis of memory to be explained. This perspective needs much more refinement it is limited to live processes unfolding moment to moment, and therefore does not provide a clear way of thinking about information that is not live, dormant, or potential. This is orthogonal to the objectives of engram biology, which seeks to explain i) a mechanism for how organisms appear to possess dormant information at all, and ii) explain how it is that engrams have the specific potential they have for eliciting a given behaviour. In other words,

how it is that the reaction of a specific ensemble, *rather than any other,* exerts such strong behavioural specificity.

4.4 — What next?

The computationalist view of information-processing and the information as relational and created view both have appealing characteristics, but neither are adequate for understanding engrams. It is unclear what form of theory would be adequate. This is not a question that would have been asked two decades ago, before the development of engram technology. It is a new problem. Nonetheless, there are several issues that any theory of engrams ought to be able to deal with: i) *arbitrariness*, ii) *coding-strength and universality*, and iii) *reading*.

4.4.i — Arbitrariness

Although decoding the meaning of any purported information in the brain is not possible if an external variable is being used as a reference point by a third person observer, this does not mean that internal codes within the nervous system, rather than between nervous system and world, are not employed. A key characteristic of a code is the use of a *symbolic* or *arbitrary* relation, such as between a triplet codon "standing for" a particular amino acid. Understanding the origins and degree of arbitrariness will be important to deciphering any internal code. For example, the taste or meaning of a chemical compound is *not* determined by the intrinsic qualities of the chemical compound but is fixed instead by anatomical wiring, namely whether the neuron expressing the relevant taste receptor is part of a labelled line to sweet or bitter cortex (Peng *et al*., 2015; Wang *et al*., 2018). Thus, to decode the meaning requires knowledge of evolutionary and developmental history of the organism. If we cannot infer anything about the properties of a chemical compound at a sensory level based on the neuronal activity it elicits, "reading out" what mnemonic features a given ensemble or complex would be capable of reinstating (without activating the ensemble and

looking at neural activity) is considerably more conceptually challenging. There is the possibility that the cells that comprise a given engram are entirely arbitrary. We want to be able to say why *these* cells, rather than *those* cells, should be involved. At least in the hippocampus, silencing the specific neurons previously labelled during conditioning while being exposed to the same conditioning paradigm simply results in a new and non-overlapping engram to form (Tanaka *et al*., 2014). Therefore an engram is constrained, but not determined by anatomy. However, if engrams are not determined by their anatomy, this does not mean that connectome-level analysis is not relevant. Instead, the kinds of invariants that may form the basis of an internal code may be structural topological invariants, rather than geometric properties of the relevant network[22]. The degree of arbitrariness may also vary regionally and phylogenetically.

4.4.ii — Coding-strength and universality

Related to arbitrariness is how strong the internal code is, namely what degree of determination over recall and ultimately behaviour a given engram is capable of exerting. Cao & Rathkopf (2019) suggest that nervous systems learn their own internal modest codes. This means that there is no one-to-one correspondence between neural signal and function. The function is indeterminate or non-disjunct. For example, Dorst *et al*. (2023) demonstrated that the activation of the same ensemble while varying the size of the environment results in different behaviours. Thus, it is unlikely there will be a one-ensemble-one-behaviour relationship. The causal contribution of a latent structure may vary as it confluences with the online contextual information of the retrieval environment (Schacter, 2007). In the case of instincts, these may involve internal codes that do not have to be learned, and may be at the stronger end of the spectrum (Ryan *et al*., 2021).

Closely tied to the concept of coding-strength is *coding-universality*, that is, whether there is any conservation in internal codes between individuals (e.g.

[22] There is an important distinction to be made here between structural topology and topological properties of high dimensional data.

how conserved one animal's internal code for a simple specific associative fear memory is relative to another's). If there is no selection pressure for there to be any level of similarity, then even if it were possible to "read out" any features based off a labelled ensemble in one animal, this knowledge would not be transferable. However, knowing the internal code may not be necessary to test this hypothesis. As discussed in Section 3.6, before the "hereditary factors" that explained the specificity to heredity were known, transformation experiments were performed. RNA transfer experiments are one kind of transformation experiment, and have been successful in *Aplysia* (Bédécarrats *et al*., 2018), but have yet to be tested in model organisms with larger nervous systems. Even if we do not think that RNA is the source of the "enduring changes" that comprise the engram in mammalian nervous systems, surrogate transformation experiments can still be performed. If a we hypothesise that the topology of the connectome is where the enduring changes are held (Ortega-de San Luis *et al*., 2023), then based off the topological characteristics of animals that have learned the behaviour, experimental manipulation that implements that same topological invariance should change the behaviour in a way that a topological theory predicts, but simple anatomy would not.

4.4.iii — Reading

Finally, any new theory of information must combat the *reading problem*. It is important to ask how these engram vehicles have their informational contents read. Here is where the standard coding metaphor can creep back in. Gallistel & King (2009) argue that the brain uses von-Neumann addressable read-write mechanisms where information is written into cells in the changes they undergo, and that it is then read by a reader mechanism that knows this code. Others deem it too awkward to apply a classical computer architecture onto the brain. Godfrey-Smith (2014) draws the distinction between two kinds of biological information. Genetic information does indeed have a read mechanism. It has a defined coding-relationship for how to go from the latent informational structure, the gene, to a protein, but no symmetrical write mechanism. Instead mutations, duplications,

and deletions occur during the course of cell division and evolution. Therefore, there is an *evolve-read* mechanism. The brain conversely has no clear reader(s), rather a *write-only* (Donahue, 2010) or rather *write-activate* system: “Evolved neural machinery has the function of introducing marks or traces into the brain as a result of experience, but these marks have useful effects on behaviour without being read” (Godfrey-Smith, 2014). While there may be a kind of distributed reading that occurs, locating the processes doing this without turning them into readers is notoriously difficult (Cao, 2012), although attempts to consider partial and multiple observers within a system have been proposed (Kolchinsky & Wolpert, 2018). Again, should it prove less cumbersome to think without the concepts of information, storage, and codes, a robust theory of how memories can be created out of accumulated changes that explains the specificity in behaviour from reactivating labelled cell ensembles is still required.

Discussion

For all the methodological advances in recent years in characterising the biological basis of memory, conceptual advances in understanding whether or not there are memory traces and what their precise functions are, or what they might look like, has not kept pace. There is a multiplicity of unclear views, areas of consideration missing, and much yet to be articulated and clarified before we have a formal theory. On a practical level, how does the above analysis help? First and most straightforward, one can ask "which sense of engram have I been using? Multiple? None of those discussed?" This piece has not aimed to be prescriptive regarding which sense of the engram should be used. In response, a classification system or *typology* could be constructed of the different mechanisms or processes involved and the different types and levels of engrams. This approach would wrestle effort away from debating what "the engram" is towards the unique reasons competing classification systems use to unify similarities and explicate differences among trace-supporting mechanisms. Additionally, it prioritises the need for work in a variety of model organisms before theorising about what traces are to be found in humans. Even if the term engram becomes reserved exclusively for the results of a given form of learning, the theory must be able to differentiate other forms of information-storage mechanisms and explain how these trace types differ. Along with putative classification systems, we also want to see a vibrant future for the contemporary systems neuroscience of memory, where different modular frameworks with clearly staked hypotheses as to which functional units (such as cells, ensembles, and complexes) best explain observed informational specificity, will compete to falsify each other in a healthily adversarial way. Our analysis also enables us to better appraise the explanatory remit of experimental approaches. We can ask "what is this study aiming to address and what do these results inform? Do they further our understanding primarily of the implementational vehicle aspect of the engram, or do they inform the more abstract, informational questions of the engram?" This piece also emphases the size and shape of the explanatory gap we face in engram biology and systems neuroscience generally, and the nature of some of the questions we will need to

engage with and move beyond in order to develop a deeper understanding. Finally, what is at stake here for engram biology and systems neuroscience? We believe it is our understanding of how the brain is capable of storing anything. Here engram biology can pursue two paths. As a sub-discipline of neuroscience coming into its own, it can attempt to help resolve the existing conceptual confusions of the wider field by articulating an engram theory within this worldview using existing, poorly defined concepts. Alternatively, the techniques and ambitions of engram neuroscience can be used as an inside job, to challenge dominant metaphors and attempt articulating a wholly different theory of what there is to be found in the brain. Importantly, if after pursuing this approach "the engram" remains even more ethereally elusive or seems to have been transformed or assimilated into different biological processes leaving in its place an assemblage of new and distinct problems to be explained, we would consider this exciting progress.

Chapter 2 — Memory neuroscience inspired AI

Abstract

The intersection of neuroscience and artificial intelligence (AI) research has a long history of reciprocal exchange. Neuroscience inspired-AI remains an important field for several reasons. The recent turn towards “scaling up” with the advent of Large Language Models (LLMs) risks overlooking potential sources of future new ideas in a AI version of the “shut up and calculate” dogma that ruled latter 20th century physics. There is still much that current AI systems struggle to do, such as continual learning, out of sample generalisation and causal physical reasoning. There is much still to be explored in bringing neuroscience into dialogue with AI research. In particular, there has been little exploration of the intersection between of contemporary systems memory neuroscience and AI. This is despite the fact that memory neuroscience and AI research share a core common challenge—storing knowledge in complex networks. This chapter seeks to establish bridges between these two fields. After establishing common ground and a brief overview of systems memory neuroscience methods, it presents core discoveries that have been made in memory neuroscience to challenge commonly held assumptions within machine learning research about neural networks and biological systems, such as challenges to the synaptic matrix model of memory storage, importance of innate priors, differences between learning in biology and artificial systems, and emerging computational models inspired by memory neuroscience. We take the connections drawn here as evidence that memory neuroscience can play and important dimension of future neuroscience-inspired AI.

Section 1 — Introduction and Motivations

1.1 — What the neuroscience of memory has in common with machine learning: a known unknown

We do not know how the brain stores anything (Poeppel & Idsari, 2022). Similarly within machine learning and artificial intelligence, we struggle to peer inside the black box of artificial neural networks to understand what and how exactly they learn. Memory neuroscience and artificial intelligence therefore share a fundamental goal: understanding how latent knowledge persists within complex networks. The kinds of answers that are commonly offered in neuroscience and machine learning are strikingly similar—typically to the effect of "information about statistical dependencies in the world is stored across the network in synaptic weights and in the overall the architecture of the network." However, such answers are overly general and when used in neuroscience and do not implicate a rigorously testable theory, for example one capable of linking neurobiological knowledge of how plasticity that occurs during learning is able to support the recall of a particular behaviour. In the case of machine learning, although we know broadly the process involved—gradient descent—it remains difficult to interpret how particular parameters of a network contribute to a given output in a "human-readable" fashion.

This is a crucial gap in our knowledge in both fields. McCulloch and Pitts' (1943) model, the first artificial neural network, was an enormously influential source of our current way of thinking about the brain, and came into existence in a cultural milieu fascinated with controlling electricity and using it to perform logical computations. The majority of our theoretical models, whether ANNs or spiking neural networks (SNNs), still today remain "spike-centric," in that their principle focus is abstracting out biological detail to create simplified models of electrochemical units. Our attempts to understand "neural computation" in neuroscience similarly involve looking for interesting features within the spiking activity of neurons and their aggregates—whether that is in vivo

electrophysiological recordings, Calcium imaging, EEG or fMRI. When it comes to understanding memory, however, a view of neural activity or computation that includes analysing only spiking is limited, because the relevant neurobiological variables that support long term memory are not maintained within recurrent spiking dynamics *per se* but rather must persist in long intervals between populations firing. They are by definition latent. Therefore, in order to address questions about memory—the nature of the stored format all our explicit and implicit knowledge comes from, our entire linguistic lexicon, all the faces we know and our motor skills—we cannot address entirely through the lens of models built on the primacy of spiking activity. This is one reason why the development of the emerging field of connectomics, the research program aimed at providing detailed "wiring diagrams" of the brains of different species, is a welcome addition. However, a connectome is static and requires interpretation. The development of optogenetic "tag-and-manipulate" methodologies in memory neuroscience (Josselyn & Tonegawa, 2020), hereafter collectivity referred to as "engram neuroscience", present a way of probing the connectome to interrogate the functionality and properties of particularly chosen latent neuronal structures.

1.2 — What is at stake

Neuroscience-inspired AI is not a singular research project with a unified goal. Attitudes differ as to what areas of neuroscience should be looked to for inspiration, and which aspect of machine learning there is translatability. One common attitude (for example in Hassabis *et al.*, 2017) is that what is needed is not to emulate biological details, but instead look for efficient algorithms the brain used for particular tasks, and bring these abstracted procedures into machine learning models. Another attitude, recognising the superior sensorimotor capabilities of animals acquired over evolution, is that future machine learning models must be embodied, in other words that AI systems should not just seek to mimic neural networks, but these artificial networks should be incorporated into artificial bodies simulations in order to undergo "embodied Turing tests" (Zador *et al.*, 2023).

We argue that in order to have neuroscience-inspired AI that is truly neuro-inspired, the insights from engram neuroscience are a crucial but overlooked ingredient. These methods have contributed a variety of core discoveries, such differentiating impaired memory accessibility from memory erasure. These and other discoveries to date challenges core assumptions, for example that spiking activity is the mechanism of all neural computation, and that information must be stored in synaptic weights. This provokes new questions for future AI systems, and a neuro-inspired AI that leaves these crucial assumptions unchallenged therefore risks not only being incomplete, but limited in offering inspiration for how to solve existing problems in artificial networks. We present these issues after providing a brief background to the tools of engram neuroscience that have made these insights possible.

1.3 — A brief background on "optogenetic tag-and-manipulate techniques"

Brain architecture can be understood to be largely genetically pre-determined with learning possible by means of plasticity mechanisms fine-tuning this architecture through exposure to new circumstances and practice (Makin & Krakauer, 2023). Some of these alterations from learning are enduring, and these enduring changes span various orders of magnitude, from changes to network architecture, to cells, and to genetic and epigenetic modifications (Josselyn, Köhler & Frankland, 2015; Ortega-de San Luis & Ryan, 2023). A natural question that follows is whether there is a level of granularity most causally relevant to the reinstatement of a given learned behaviour. Even though it remains unclear whether discrete memory traces (or "engrams") have been or could ever be identified, the idea itself drives an important systems neuroscience research programme aimed at understanding the biological basis of memory. While the neuroscience of learning and memory often involves characterising plasticity mechanisms that correlate with or are necessary for memory, engram neuroscience uses a different set of techniques to label neurons during encoding and reactivate or manipulate those same neurons during recall. The technologies employed typically involve using transgenic

animal models to label neurons that are activated during a window in which associative learning occurs, and reactivating or inhibiting those cell populations with light or chemical stimulation at a later timepoint that tests the necessity and sufficiency of these labelled ensembles (Reijmers *et al*., 2007, Liu *et al*., 2012). There is overlap between those cells active during the learning window and at the time of recall (% varies by region of the hippocampus), and artificially inhibiting the cells tagged during learning at the time of recall impairs expression of the behaviour, while artificially stimulating engram cells is enough to cause freezing behaviour in new contexts different to training contexts, and stimulation can be used to create artificial associations in the absence of direct experiences (Josselyn & Tonegawa, 2020). Those cells, ensembles, and pathways that are necessary and sufficient for that particular behaviour are considered "engram cells/ensembles/ pathways," in contrast to randomly tagged control "non-engram cells," which when activated do not produce the learned behaviour. Those cells that can become engram cells can be inhibitory and may also include non-neuronal glia, and form linked complexes distributed across the brain within which different cells play different roles during encoding and recall across the lifespan, (Guskjolen & Cembrowski, 2023). In short, these techniques go beyond mere optogenetic or chemogenetic activation or inhibition of neurons, they probe specific subsets of cells functional at specific times, making the technology a crucial means of going from brain structure to brain function.

Section 2 — What we have learned

2.1 — From classical to an alternative view of memory

2.1.1 — Connectivity

The dominant view, postulated by Hebb (1949) and seemingly validated by the discovery of long-term potentiation (Bliss & Lomo, 1973; Lynch, 1998), along with mechanisms such as spike-timing dependent plasticity, cellular consolidation and behavioural timescale plasticity (McGaugh 2000; Kandel, 2001; Bittner *et*

al., 2017) is that changes in synaptic strength explain the formation of memories. The updating of acquired knowledge (or "learned representations") in this view amounts to changes in strengths of existing synaptic connections, and that memory traces are somehow stored across synapses. However, there are studies that have shown that changes in synaptic weight are not necessarily required for learning (Chen *et al*., 2014; Ryan *et al*., 2015).

Therefore, the view that memories are encoded in the scalar values of a weight matrix is likely insufficient in the case of real neural networks. What other mechanisms can support learning if not changes in synaptic weight? This realisation shifts emphasis onto other candidate plasticity processes, including (but not limited to) another mechanism of synaptic plasticity—structural plasticity due to changes in synaptic wiring (e.g. due to synaptogenesis, axon and dendritic growth or elimination), a mechanism easily confused for plasticity of synaptic weight (Chklovskii, Mel & Svoboda, 2004). In sparse networks, this form of plasticity supports the formation of many functionally distinct circuits for encoding learned information through the flexibility of a post-synaptic unit's combinations with a large set of pre-synaptic units (ibid).

It has been shown that learning modifies the connectivity pattern between specific engram-ensembles (Vetere *et al.*, 2017), and that the updating of prior associations in light of new evidence involves the growth of new connections between distinct specialised regions to reroute future activity (Ortega-de San Luis *et al.*, 2023). This supports the hypothesis that the mechanism which explains how prior knowledge is updated in a way that persists long term is through the modification of connections between unique ensembles—rather than or in tandem with—changes in the strengths of existing synapses (Ryan *et al*., 2021). Memory traces therefore cannot be understood as the values in a weight matrix, because the connections that comprise the matrix itself are constantly shifting.

In contrast, the field of artificial intelligence has largely assumed that one must hand-craft a pre-specified network architecture, and learning is the modification of the weights between already connected units in a tabula rasa network. Again,

McCulloch & Pitt's 1943 model may have had a "founder effect" in establishing particular view of actual neural systems. In later connectionist models, still descendants of McCulloch and Pitts' original model, what remains when there is no "activation" is typically understood to be distributed across the network in specific the weight and bias values. However, in these models these are the only free parameters left to vary when an architecture is set, so it would be surprising if everything that needs to be accounted for under the construct of "memory" was to be described any other way. These values are difficult to interpret, and it is possible that they are too coarse a description-level to be useful for understanding memory in biological networks.

What evolution can work on are neuronal wiring rules and patterning, which Zador (2019) argues makes network topology/architecture fundamental considerations in any neuro-inspired AI. Indeed, visual receptive fields loosely inspired the use of highly constrained wiring used for CNNs (ibid). Although it is clear that there is degeneracy at the level of neural circuits for implementing certain computations, such as different algorithms and neural implementation for sound localisation in barn owls versus guinea pigs (Grothe, 2003), Langdon, Genkin & Engel (2023) have highlighted how computation over low-dimensional manifolds can be linked to unique circuit architectures, for example certain recurrent architectures can support categorical decision-making, others wiring motifs enable computing an animal's heading direction.

It is clear that even after developmental sculpting of neural circuits, during the lifetime many different mechanisms capable of influencing memory formation converge on changing the connectome, including creation and retraction of synapses, synapse elimination by phagocytic microglia (Wang *et al.*, 2020) and astrocytes and oligodendrocytes (Kol & Goshen, 2021), and neurogenesis (Akers *et al.*, 2014). It remains an exciting open question at what level of granularity we ought to care about changes in structural topology for understanding learning and memory, and how these alterations are capable of sculpting network population dynamics. Nonetheless, the picture that emerges is one distinctly at odds with artificial networks that maintain fixed architectures. Although there have been

steps towards “anatomically constrained ANNs” (see also Saxe, Nelli & Summerfield, 2021), architectural changes are limited to the level of modifying artificial synaptic weights. A second core problem with the synaptic weight view of memory is that it does not allow for a mechanism of acquired knowledge that is innate.

2.1.2 — Innate priors

A fundamental challenge for deep learning is building prior knowledge into untrained networks. Marcus (2018a) gives the example of how the fact that “tabula rasa” networks can approximate Newton’s Laws of Motion from pixel-level data, and be trained to play Go (Marcus, 2018b), can give the impression that building prior knowledge into ANNs is simply unnecessary. However, given that issues such as poor generalisability beyond training sets, unrealistic levels of training examples and enormous energy expenditures continue to plague ANNs—while biological brains have evolved over millions of years to come both “hardwired” and “prewired” (Marcus, 2004) and are far more adept at generalisation, do minimal supervised learning, and run on a few Watts—the emerging field of neuro-AI has emphasised paying more attention to innate knowledge. Zador (2019) has even argued that unique yet-to-be-discovered unsupervised learning algorithms from biological systems are unlikely to be the answer to why animals are capable of few-shot learning. Simply looking to brain for algorithms is not enough. Instead, brains come with numerous kinds of innate “scaffolds,” structures that support e.g. navigation or the ability to recognise faces, and what is learned is only possible because of its being built off these systems, including memory of specific locations and faces (ibid).

Presently, a typical ANN starts with a blank architecture with randomly initialised weight values. The hope is that, at least in sufficiently large networks, sometimes there may be random initial configurations that are useful for the task the network is about to be trained to perform. However, there is no reliable way to “bake in”

these specific priors into the network before training—we cannot design what these precise priors should be and where they should go.

In biology, one emerging hypothesis (Ryan *et al.*, 2021) states that there is no difference, at least in terms of biological substrate, between a hypothetical "innate ability system" and a "learning system". The same optogenetic and chemogenetic tools used in studies of learning in engram cells have also been used for the tracing and artificial manipulation of hard-wired "labelled line neural circuits" for innate behaviours, such as for taste (Wang *et al*., 2018), olfaction (Root *et al*., 2014), mating (Anderson, 2016; Ishii *et al*., 2017) and predation (Han *et al*., 2017; Shang *et al*., 2019). These "innate engram" studies are important because they provide further corroboration for the idea that innate priors are located within the topography of the connectome. Knowledge acquired during the lifecycle and instincts acquired through evolutionary time, even if they arise out of different processes (the growth of the brain and formation of connections during development, and learning through plasticity modification of existing architecture) may therefore be structurally isomorphic (indistinguishable) at the level of the brain, meaning brain networks themselves may not differentiate between whether a memory is learned or innate (Ryan *et al*., 2021).

2.2 — Erasure and inaccessibility are different

One of the most suggestive and consistent discoveries using engram techniques is that learned knowledge appears to be highly resilient, persisting even when previously assumed permanently lost in various types of amnesia (Ryan & Frankland, 2022). In other words, forgetting does not always necessarily imply erasure of learned knowledge, rather memory can survive but becomes inaccessible. This has led to the conjecture that the transition from latent knowledge from accessible to inaccessible occurs through changes within or in the neighbourhood of the specific cell ensembles relevant for that learned behaviour (ibid). Amnesia may therefore be understood as a behavioural phenotype that also occurs either whenever changes alter the probability of the

relevant ensembles firing given the presence of natural retrieval cues, leading to a retrieval deficit, while severe damage to these ensembles causes amnesia via storage deficits. The application of these new genetic techniques has made possible dissociating between an inability to retrieve a memory due to complete loss of memory (erasure), or an inability to access, in a variety of cases. In cases of infantile amnesia, associations formed during infancy that are behaviourally absent in adolescence can be artificially recovered by stimulating those cells that were labelled during learning (Guskjolen *et al*., 2018; Power *et al*., 2022). The same has been observed in retrograde amnesia induced by protein synthesis inhibitors (Ryan *et al*., 2015), models of early Alzheimer's (Roy *et al*., 2016; Perusini *et al*., 2017; Poll *et al.*, 2020), sleep deprivation (Bolsius *et al*., 2023), retroactive interference (Autore et al., 2023), natural forgetting (O'Leary *et al.*, 2023), and after traumatic brain injury (Chapman *et al*., 2024). This challenges existing views about what occurs during learning and what it means for latent knowledge to persist in the brain. In many cases where researchers thought inducing amnesia meant erasure of memory traces, these memory deficits can instead be reversed with appropriate retrieval cues, shifting emphasis to the factors that influence retrieval processes (Alfei *et al*., 2024).

These shifts in perspective within memory neuroscience also open up numerous questions about the differences between biological and artificial networks. Both brains and ANNs are said to face the "stability-plasticity dilemma" (Mermillod, Bugaiska & Bonin, 2013), where too much plasticity for new learning causes excessive forgetting, while too much stability challenges new learning. In the extreme cases of catastrophic forgetting in ANNs, what was previously learned is interpreted to be permanently erased when new learning interferes. In contrast, on top of all of the innate knowledge that comes pre-wired in biological brains as discussed in Section 2.1.2, it appears that knowledge created during the lifecycle is capable of persisting even in cases of interference and apparent memory loss, meaning existing memory traces do not have to necessarily be erased during the acquisition of new knowledge, suggesting a dissociation and mechanisms through which they persist but their accessibility state becomes up or down-weighted.

In summary, the studies discussed in this section have demonstrated, at least in principle in biological brains, that memory absence and permanent memory erasure can be partitioned. This partitioning may be a mechanism that enables real neural networks to overcome brittleness/overfitting one one hand and catastrophic forgetting on the other—interactions with the environment vary the accessibility of knowledge relevant for the present over a network capable of carrying forward all previously learned knowledge in a preserved but inaccessible state. Whether or not similar properties exist in ANNs will be discussed in Section 3.3.

2.3 — Memory is embodied (can be non-neuronal)

Several important discoveries have also called into question whether all memory traces must exclusively be neuronal. For example, using the same kind of genetic tag-and-manipulate techniques discussed above, Koren *et al*. (2021) showed that the insula stores traces of specific immune reactions, and the reactivation of those neurons encoding a specific immune response is sufficient to induce a similar immune response in the body even in the absence of the original insult. This led Koren & Rolls (2022) to claim that the components comprising a given memory trace of an immune reaction are distributed between specialised brain networks and in the previously inflamed tissue (in the form of receptor expression profiles and memory immune cells). In other words, the neuronal contributions to certain kinds of memory traces are necessary but not sufficient, and are integrated with persistent changes in the body. This also means that there are specific ensembles in the brain for particular reactions in the body, and that it is not a phenomenon of organism-wide habitation.

The brain's own immune cells have also been shown to control whether forgetting occurs (or does not occur) through altering the connectome. Chen *et al.* (2024) demonstrated that microglia are responsible for the extinction of particular stress responses through directly engulfing the dendritic spines of inhibitory neurons activated by the acute stress, while Power *et al.* (2023) showed that immune activation during pregnancy prevented infantile amnesia occurring in male mice

offspring, implicating a crucial immune signalling systems in mediating the effect. When it comes to the consolidation and expression of memories, other systemic signalling systems are capable of playing important roles. Glucocorticoids released from the adrenal glads in response to stress differentially effect retrieval of aversive memories and fear-extinction memories (de Quervian, Schwabe & Roozendaal, 2017; de Kloet & Joëls, 2023), and recent versus remote memories (Brosens *et al*., 2023). Together, studies such as these underline how, in biological brains, processes of learning and forgetting are part of general bodily homeostasis, and overlap with and are integrated into the same systems responsible for internal damage detection, repair, and stress regulation. On longer timescales, this work underscores an important difference between brains to ANNs: biological neural networks evolved within bodies to sense the viscera and effect responses throughout the body in turn. In Section 3, we attempt to explicate some consequences of these insights from memory neuroscience for current and future machine learning design and implementation.

Section 3 — How this is relevant for AI

3.1 — New theories of forgetting

When information is lost in ANNs due to training interference, it is presumed to be permanently erased. However, as discussed in Section 2.2, learned associations previously believed to be irreversibly destroyed—for example infantile amnesia from the sheer amount of synaptic remodelling that brain networks undergo during early post-natal development (Guskjolen *et al*., 2018; Power *et al*., 2022), or during the damage done by neurodegenerative conditions (Roy et al., 2016; Perusini *et al*., 2017; Poll *et al.*, 2020)—do persist but are “inaccessible” and are only retrieved through artificially induced selective activation. This introduces another dimension to the many plasticity mechanisms: short-term plasticity (facilitation, augmentation, fatigue etc.) and long term plasticity (long-term potentiation and depression). Memories do not simply traffic from and to short

term memory to long term (recent) to long term (remote), and if they are not retrievable under certain conditions this does not automatically imply that they have been lost from long term memory.

An important question this prompts is whether there exists a comparable dissociation between (in)accessibility and presence/absence of learned knowledge in ANNs, or whether this is a property of biological networks currently absent with no clear analog in ANNs. Let the set of all useable domains of the biological network that are accessible (as opposed to inaccessible) be referred to as B_a, and B_i be the remaining set of all inaccessible states, regardless of their reason for being inaccessible. What might B_a correspond to in ANNs? And if there is a separation, what algorithmic principle(s) might govern traffic between accessible and inaccessible states, $B_a \rightleftharpoons B_i$? It should be noted that neuroscience does not yet offer a theory of how this works in biological brains, it being unclear whether it is modulated by late-phase long-term potentiation, excitability state, neuromodulators, or competition between ensembles. What follows in Section 3.1 is therefore conjecture that attempts to establish potential new connections between memory neuroscience and AI.

Regarding the first question of whether there is a comparable analog to the modulation of accessibility in ANNs, there are several comparisons that could be suggested. One possibility is that the set and contents of what are accessible at a given time corresponds to whatever is within the context window of a large language model (LLM). The context window is unlikely to be analogous to human working memory, which can only hold a few items at a time. Present context windows can hold up to tens of thousands of tokens, but is this really comparable to the sum of all available knowledge available for inference in a given biological system, and if so, which—*C. elegans*, a fruit fly, or a rodent?

A first objection might point out that information within the context window is not stored in network weights but in activations, and whether or not a component of a biological network is accessible or not is dissociated from whether or not it is

"firing" during a given interval. On these grounds then, a more intuitive comparison might be that the context window instead corresponds most to the biological changes of short term plasticity.

In biological networks, attention modifies the firing properties of neurons (and whether there are likely to be any downstream intracellular changes as a result) through the use of neurotransmitters, but plasticity mechanisms through a variety of other processes (enhancement or depression) can also modify the relative importance of elements of the network without necessarily permanently altering the weighting of synapses.[23] Instead of the context window, this form of short term plasticity may also be more analogous to updated input in Retrieval Augmented Generation (RAG) LLM architectures. Here, a custom up to date database is converted to external embeddings capable of influencing the overall activation of the without changing structural features of the LLM (Lewis *et al*., 2020). However, the external databases are designed by human engineers and with specific types of human-created data that is amenable to being converted to vector embeddings. The additional information in the external memory is not incorporated into the network. In conclusion, it appears that there is no strong candidate analog of a dissociation between accessibility and presence. This presents an exciting opportunity for the development of biologically-informed learning theories.

3.2 — Evolution and learning are not optimisation & a proposed "neutral theory of plasticity"

[23] However, some caution is warranted when attempting to draw literal comparisons between biological and artificial networks. There are fundamental differences. ANNs do not have inhibitory neurons, complex arborisations, or large within-network variation in unit-type. Given that it remains unclear whether ANNs are models that map neatly on to physical organisation or are instead coarse-grained approximations that are not implementationally isomorphic (e.g. nodes not really being neuron-like, but instead in capture some higher-level properties of groups of neurons under certain circumstances), it may be possible to relax the drawing of comparisons based on implementation and focusing instead on general properties. If not committed to maximising structural isomorphism, then the fact that the context window deals only with activations may not be an issue. If biological plausibility is preferred, then a better analog for modulating accessibility may be a semi-permanent dropout mechanism for unique modules of a network (as opposed to singular units).

In ANNs, there is no evolution-learning distinction. To frame biological systems in similar terms, we might say they have undergone millions of years of extracting statistical regularities, with selection pressures over intergenerational time providing "reinforcement" and evolution acting on the genome, not by encoding representations and cost functions, but by modifying wiring rules and patterning during brain development (Zador, 2019).[24] Treating evolution and learning as a continuum might allow more accurate comparison between organisms and ANNs.

However, setting evolution and learning up in a way that they can be collapsed into a process comparable to the training of ANNs overlooks important differences. Perhaps most crucial is the often forgotten fact that *evolution is not optimisation*. There is no landscape with a global optimum which all organisms are agentially striving towards. This is the old and long-refuted hypothesis of directed evolution or "orthogenesis," which exists due to the unfortunate conflation between evolution and "process" outside of biology and in particular within the technology industry.

Given the centrality of defining and optimising objective functions, the AI community may be mistaken for thinking of biological evolution as the optimisation of "fitness" through a selection process. This is because natural selection is only one means of evolution. At a molecular level, the majority of evolution occurring within a population follows a process of random drift which is either neutral (Kimura, 1983) or nearly neutral (Ohta, 1973). Evolution, therefore, is a process both highly "open-ended" as well as functionally constrained. Rather than solely evolving to only ever "directly fit to nature" through adaptation, we get the evolution of *evolvability* itself—the ability to generate potentially useful and non-lethal phenotypically novel traits (Kirschner & Gerhart, 1998). This means many aspects of molecular biology are left "free to vary" over evolutionary time, allowing for the exploration and development of novelty that, in some cases, may later become functional (Koonin, 2016). A consequence of this is that not

[24] Kaznatcheev & Kording (2022) argue for a direct analogy between evolution discovering developmental deconstraints and cultural evolution of deep learning advancements. However, because this is development of machine learning models at the hands of human engineers rather than the networks themselves, it is not a direct comparison.

everything in the molecular world—such as a genomic sequence—is necessarily "for" anything. It is not the result of small adjustments made during generational training epochs towards some high-dimensional optima.[25] Another consequence is that when evolutionary pressure is strong it can prevent exploration of novelty (ibid). Lastly, maximum fitness is not equitable with "ground truth" and optimisation of fitness does not always lead to "success." Species may be seen to "optimise" for a particular trait even when it is maladaptive, as in classical Fisherian runaway observed in sexual selection of many species (Fisher, 1930) and evolutionary-to-extinction (Lehtinin, 2021).

We argue that it is worthwhile conceptualising learning in similar terms to evolution.[26] In other words, to have a "neutral theory of plasticity/learning" as opposed to an "objective-function-only view of plasticity/learning."[27] First, a neutral theory of learning would posit that although it can be, not all learning is goal-directed, and that it is not the case that the various forms of plasticity exclusively implement a form of biologically approximated gradient descent. Thus, not only are biological networks decidedly not tabula rasa networks, but there is also reason to question the assumption that during a given lifetime they respond merely in lockstep to only ever approximate "invariances in the world" that could even in principle be "fully learned" with unsupervised learning rules like coincidence detection with infinite time. In other words, learning is not about gaining better access to reality, but simply whatever allows for perpetuation proliferation in a particular niche, which may include social worlds of arbitrarily fixed conventions at a particular time and not in for the future. This cuts against the direction of the dominant ways of thinking in machine learning and the

---

[25] These insights may be counterintuitive to those coming to biology from an engineering background, where the *telos* or purposes of the system are created by humans and imposed on through design principles. To the extent that there exists a grand multivariable optimisation equation, it would be one where cost-functions comprising it are constantly modified and sift in and out of the general equation, meaning the latter cannot in principle be defined.

[26] This is far from an original connection to make, and there are many ways to frame it. For example an example drawing the parallels between Bayesian inference and fitness optimisation, see Czégel *et al*. (2022).

[27] Exemplars of this latter approach include many ideas emerging at the intersection of biology and machine learning. For example, Hasson, Nastase & Goldstein's (2020) argument that evolution and learning is a "direct fit" to nature, and Richards & Koerding's (2023) argument that plasticity "has always been all about gradients".

disciplines that it has influenced, including computational neuroscience. It has come under only select criticism from within machine learning, for example by Stanley & Lehman (2015), who argue that it is precisely the lack of open-endedness imposed by defining objectives that prevents creativity in AI, whereas evolution throughout the eons and culturally is characterised by open-endedness (ibid; Kauffman, 2019).

It is possible to interpret existing literature in terms of a more "neutral theory of learning." Given neurogenesis (Akers *et al*., 2014) and the rapid turnover of spines within the hippocampus (Attardo, Fitzgerald & Schnitzer, 2015) the synaptic "ground" of biological networks important for learning appears to be dynamically shifting. This is especially clear from the the accumulating evidence for the drifting coding properties of neuronal populations (Ziv *et al*., 2013; Rule, O'Leary & Harvey, 2019). Although the level of drift varies between brain regions and in some cases is very tightly constrained, it has been shown that these drifting codes do not have to pose a threat to previously learned tasks or associations, as feedback mechanisms can ensure stable read-out (Rule and O'Leary, 2022). Rather, neural populations are free to vary as long as this variance is kept to off-task dimensions. It appears that, much like evolution, open-endedness at a lower level of organisation is combined with constraint at the phenotypic (e.g. behavioural/task) level. There are a variety of non-mutually exclusive hypotheses for why drift may occur and be a generally adaptive process, such as promoting exploration and preventing overfitting (Kappell *et al*., 2023; Aitken *et al.*, 2022; Ratzon, Derdikman & Barak, 2024). We believe that persevering a portion of all plasticity for neutral exploration of the connectomic space may be beneficial for generating potentially useful population codes (Wolf *et al.*, 2017), regardless of whether this is part of an active ongoing consolidation process or occurs in the absence of learning, and that this may be a potential function of so-called "silent plasticity." The real world is not divided up neatly into natural kinds of tasks or benchmarks. Indeed, there are times when it is not always clear what specific tasks are being performed, or if there is a specific goal in mind when engaged in certain activities. Learning therefore adapts and repurposes what we already have, we do not learn what we learn for specific reasons, since those cannot be stated

ahead of time. In other words, just because memory/priors/learned representations can be used to inform goal-directed task-specific decision making, does not mean learning is always goal-directed and that there is an identifiable function for all memories.[28]

To summarise with respect to previous sections, biological brains may possess the ability to continuously form connections that may not be useful for a specific purpose or goal (but still follow the same plasticity mechanisms of explicit learning and are not erased entirely when unused). Associations can be made which may not be "for" anything (and may never be). Connections that are not for purposes relevant at a given time are not purged/erased, but rather may become silent or inaccessible, only to later be reawakened and made accessible if relevant (Vardalaki, Chung & Harnett, 2022; Ryan *et al*., 2021).[29] Most abstractly, the shift in worldview is that the point of all plasticity is not to approximate with high fidelity and redundancy regularities of an external world, but create latent possibilities that may happen to be useful later in un-anticipatable environments. A naturally dynamic connectome which combines plasticity processes capable of neutral exploration with goal-directed learning would serve as a substrate capable of parallelising exploration and exploitation.[30] How these latent motifs in the connectome may be further understood from a computational perspective will be discussed in Section 3.4.

3.3 — Functionalist and implementational computational biological models

---

[28] In contrast to both evolution and learning, *development* does have stereotyped destinations that the system actively tries to reach/must satisfice. A relatively high level of fidelity is therefore required and is the reason developmental genes are tightly regulated to avoid foetal non-viability and cancer post-natally.

[29] Forming associations that do not (yet) have a clear use/function is not necessarily the same as forming spurious associations, which when chronically present would be considered a pathology (e.g. psychosis).

[30] And may be construed as implementing a "parallel terraced scan" algorithm (Rehling & Hofstadter, 1997).

The stark differences in capability between machine and biological learning provides one impetus for building more biologically plausible models. What would it take to build a system that does continuous learning and does so without catastrophic forgetting/interference? There are two differing (and complementary) approaches. One seeks to combine with what we have learned from biology at a functionalist level with ANN modelling techniques, and designs models that capture an important abstract property of biological memory systems, without being committed to building a simulation out of biologically plausible units (Hassabis *et al.*, 2017;). The other contends that implementational realism should be prioritised (Zador 2019; Hiesinger, 2021) and biological substrate should be mimicked as close as possible or as is computationally tractable. This extends to SNN simulations as well as building neuromorphic architectures that attempt to embody biological circuitry. We discuss existing attempts at both in the context of memory research.

3.3.1 — Functionalist models

Computational modelling of memory systems has an established history. For example, the complementary learning systems (CLS) theory (McClelland, McNaughton & O'Reilly, 1995; Kumaran, Hassabis & McClelland, 2016, with roots in Marr (1971)) aims to explain computationally *why* we have the partitioned architectures of the hippocampus and neocortex at all. One system (the neocortex) is specialised for gradual learning of abstract hierarchically structured knowledge in a given domain (language, skilled action, perception, etc), but which is consequently poor at few- and one-shot learning, and which if attempted leads to catastrophic interference/forgetting (Kumaran, Hassabis & McClelland, 2016). To complement this system, another (temporal lobe structures including the hippocampus) specialised for rapid learning of unique instances in detail (using a pattern separated, non-similarity-based coding scheme), which came at the expense of generalisation (ibid). By optimising for different parameters, these two distinct learning systems are differentially recruited and can work together depending on task demands.

Sun *et al.* (2023) is a recent example of a computational model in the CLS vein. They used a Shallow Linear Network as a fully interpretable network model (for the neocortex) connected to a symmetrically weighted associative network (to model the hippocampus) to understand when and how much "replay" for generalisation (from the hippocampal analog into the neocortical analog) is optimal in different environments. Importantly, here shallow linear networks are not understood as literal representations of the neocortex, nor does the hippocampus look like a Hopfield network. Instead, this work continues to demonstrate that using highly abstracted models is still a fruitful avenue that continues to generate insights into memory systems.

There is also much to be gained from increasing the level of realism of existing ANN models. Chandra *et al*. (2024) presents a synthesis of previous models attempting to capture different elements of the hippocampal-entorhinal system, including its involvement in navigation. They found that by building a model that included hard-wired grid cell analogs, their modified continuous attractor network could overcome the phenomenon of catastrophic forgetting, something that has long plagued various versions of attractor networks (ibid). They achieved this through a variety of means, including separating the fixed-point attractors corresponding to discrete memory items, from the content to be stored, creating a "memory scaffold" of pre-structured states with no content (ibid). This fits with views of the hippocampus as not storing the all of the content of memory *per se*, bur rather a highly compressed set of "pointers" or "links" needed to reactivate the appropriate populations of the neocortex (McNaughton, 2009; Goode *et al*., 2020).

A memory can refer to both in the general ability for remembering as well as to singular instances, such as episodes or facts. If we want to explain at the most fine grained instance, then we may have to consider the precise population structures of groups of cells (both identities and order of activation) responsible. CLS theory is separate from but shares some similarities with the theory of systems consolidation (Wiltgen and Tanaka, 2013; Squire *et al*., 2015), which proposes

that memories are first formed in the hippocampus and eventually later in the neocortex, where they become hippocampus independent (although not always). Engram optogenetics studies have played a crucial role in the development and refinement of systems consolidation theory. For example, Kitamura et al. (2017) showed that hippocampal engram cells for a particular fearful experience become silenced over time, while engram cells in the amygdala were necessary for the fearful experience to be maintained, while further engram studies have demonstrated the role of the prefrontal cortex in consolidation (Teranova *et al*., 2023). The question this research poses is how best to incorporate these further biological details into computational models.

3.3.2 — Implementational models

Tome *et al.* (2024) is the only attempt to build an implementational model explicitly informed by recent memory trace research. They built a spiking neural network (SNN) model of LIF units connected in a way to mimic the architecture of the hippocampus as closely as possible. Crucially, as is currently impossible in ANNs, they included inhibitory units. The lack of inhibition in ANNs is often glossed over because of their successes, but inhibition is an intrinsic keystone of brain function, serving many purposes including massively expanding the computational capacity of a real neural networks through effectively transiently changing the functional connectome (Buzsaki, 2007). Their model predicted dynamic cell identity within the population of cells responsible for storing a given a “memory” and highlighted the essential role for inhibition in selecting the appropriate trace at recall and inhibitory plasticity in consolidating memories with sufficient selectivity (Tome *et al*., 2024). Further studies are needed, and while SNNs generally are attempts to create artificial models that feature more biologically realistic units than classical ANN activation units, but because of their greater complexity and a lack of investment comparable to ANN theory and the optimising their use for running on GPUs, SNNs remain more computationally

expensive to run on digital computers (Furber, 2016).[31] This is one driver for building neuromorphic hardware capable of running SNNs at a fraction of the cost (Jaeger, Noheda & van der Wiel, 2023).

An important question to ask in the context of of both abstract and implementational models is "how should we understand what a memory trace should be in one of these models?" —and whether or not this is a sensible question. Even SNNs and neuromorphic chips, in being based on electrophysiological spiking activity, can only go so far in terms of biological realism. Neurotransmitters and hormones (Mei, Muller & Ramaswamy, 2022), viruses-like RNA capsids (Pastuzyn *et al*., 2018) and all three other glial cell types (astrocytes, oligodendrocytes, and microglia) have all been implicated in memory processes (e.g. Kol & Goshen, 2021; Steadman *et al*., 2018; Chen *et al*. 2024). Although to exert an effect behaviour all biological processes must eventually converge on exerting some change to electrophysiological firing, alterations to memory do not need to converge on activity, they can influence behaviour through structural plasticity changes, processes currently omitted from computational models.

3.4 — Gaps in computational theories of memory

In the previous section we talked about simulating biological networks on computers or mimicking certain properties in neuromorphic circuits. Now we switch to discussing computation in living systems. As a poorly defined and multivalent term, there are various ways to think about memory from a computational perspective. We break these down according to why-, how-, and what- questions. We argue that there are as yet no "what-formalisms" for understanding memory computationally.

[31] But this research is continually improving and continues to provide computational explanations for the efficiency of real neural networks (Perez-Nieves *et al*., 2021).

“Why have memory?” is straightforward to answer from a computational perspective. All organisms have evolved the ability perform automatic online-processing through entrainment with the relevant inputs of their respective niches. To modify the system responsible for generating and maintaining online dynamics is costly, but storing the results of past computations (“amortised inference” (Gershman & Goodman, 2014) is a worthwhile investment (Dasgupta & Gershman, 2021). In the wild, the relevant selection pressure may occur when the costs of learning and preserving what is stored outweigh the cost of maintaining grey matter for the lifespan of the organism, e.g. despite seasonal fluctuations in food availability that require location memory for storing food (Sherry & Hoshooley, 2010).

For how and what models of memory, we borrow from Jaeger’s (2021) framework for exploring generalised theories of computing. In this framework, how-formalisms describe procedures while what-formalisms are human-interpretable formal objects. For example, in digital computing a what-formalism would be a symbolic logic, while a how-formalism would be a programming language. In probabilistic computing, a what-formalism would be a probability distribution, and a how-formalism would be a sampling algorithm. In dynamical computing, a what-formalism may be a better way to understand geometric features like fixed points and attractors, while the how-formalism would be ODEs (ibid). When it comes to understanding memory, it seems there are some existing general how-formalisms. These are models that use particular network architectures and specific update rules to examine how a network evolves over time, for example existing associative and energy-based NNs. However, there is more work to be done, such as in understanding the algorithms of forgetting. There is no external scanner that sifts through the connectome and decides which ensembles should be silenced, which potentially implies a mechanism of inter-ensemble competition. This prompts the question of what local rule is followed/what the criteria are for determining which ensembles are expressed and which are silenced.

Ultimately, what are the objects what we would describe as being manipulated here? It seems that there are no what-formalisms for modelling memory systems.

One way to do this may be to consider a symbolic approach to modelling memory. We next turn to consider potential issues when it comes to describing memory symbolically within the context of existing neurosymbolic AI discourse.

3.5 — Neurosymbolic AI and Memory

What counts as neurosymbolic-AI varies according to different heuristics, examples vary from LLMs (because their inputs and outputs are text tokens), AlphaGo (which uses Monte-Carlo tree search) to the Neural Concept Learner (Mao *et al*., 2019). This domain of cognitive science research is expressly interested in "higher cognition" and as a way to understand the difference between humans and primates and ANNs (e.g. Sable-Meyer et al., 2022). The distinction between two reasoning systems popularised in Kahneman (2011), between one fast and implicit ("System 1") and another slower and deliberative ("System 2"), has been observed to analogously track the two broadly different historical approaches to AI—with the more statistical, black-boxed approach of artificial neural networks and machine learning corresponding to System 1, and classical symbol manipulation corresponding to System 2. But are these symbolic frameworks only appropriate to use when trying to model the abstract syllogistic reasoning that humans are capable of? Given that the ability to remember is not uniquely human but is present across the tree of life, does this mean there is no scope to consider memory symbolically? We take as broad a view as possible of what may count as neurosymbolic AI, and contend that we need not restrict describing computation in terms of abstract symbolic operations exclusively to overt reasoning.

Kazanina and Poeppel (2023) present a case that the hippocampal-entorhinal system that forms the rodent navigation system can satisfy the requirements of a "language-of-thought" framework (Quilty-Dunn, Porot & Mandelbaum, 2022). They argue that the selective firing properties of place cells, border cells, and

other cell-types, are sufficiently abstract to function as predicates that can freely couple and uncouple to a variety of input arguments, while connectives could be constructed out of “conjunctive cells” e.g. border and direction combination-selective cells (ibid). They argue that these properties imply the hippocampal entorhinal system does operations more abstract that merely forming simple associations. They also include the crucial clarification that this interpretation does not imply that all thought is spatial or hippocampal. Additionally, just because an LoT framework can be used to describe the rodent hippocampal-entorhinal system does not necessarily imply that rodents are capable of System 2 reasoning. It therefore demonstrates that neurosymbolic AI does not always have to be thought of in System 1 & 2 terms.

In a different vein, one associated more with symbolic thinking within biology, Dragoi (2024) has proposed that arbitrarily pre-configured ensembles of place-cells are generated in sequences that provide a "syntax" that can later be "selected" during learning, although the exact mechanism that determines how these groupings are assembled remains unknown. While Kazanina and Poeppel (2023) provide an attempt to develop a more symbolic way of thinking about the hippocampal entorhinal system in terms of the firing-selectivity of specific cell types, and Dragoi (2024) does this for sequences of cells in time, there is also the possibility of attempting to do this for physical structures within brain networks.

Granting that changes that occur during learning occur in context, namely within a living tissue shaped by millions of years of evolutionary processes, its unique developmental and personal history (Ghazanfar & Gomez-Marin, 2024), it is still worth asking whether there are universal motif changes to a given network that always implement the same alteration in a way that they may be understood as an abstract primitive operation. In other words, given a network, if we always created ***these*** new connection changes and synaptic strength adjustments in a variety of animals, would we always create, say, a generalisation from the most recent context, or a memory of a place, or a memory that interferes more or less with previously acquired knowledge, for example?

While the functionality/logic of a given ensemble is relational in that it depends on what it is linked with, lines must be drawn at some point. In the coarsest case, a what-formalism should at least capture something about the differences between hippocampal and cortical ensembles. In the spirit of Kazanina and Poeppel (2023), for example, the memory objects of the hippocampus may be understood as bundles of logical connectives, with cortical ensembles their referent. This would go a way to establishing a kind of mid-level to use symbolic descriptions between the classical System 1/2 dichotomy.

Although there has been much development in terms of applying topological analysis to data (to find shape in data), developing tools to analyse the topological properties biological networks is a different task, and one that has only partially been explored (e.g. Stolz *et al.*, 2019; Benjamin *et al*., 2023; Sen *et al.*, 2022). Ultimately, this is an as-yet unexplored avenue of collaboration we would be very excited to see happen. Having a greater understanding of the properties of particular network topologies would may also assist in the field of Interpretable AI (XAI), which seeks to makes ANNs less like black-box and more transparent and human readable

Concluding Remarks

Memory is an interesting edge case to consider when it comes to building computational models, because the majority of our theoretical models abstract out biological detail in order to focus on modelling electrophysiological units, leaving other aspects of biology crucial for memory excluded or demoted in consideration. While there is much that remains to be done in the neuroscience of memory, both experimentally and in terms of the development of new theoretical frameworks, we have demonstrated that there is a crucial place for the insights discovered from recent memory neuroscience for future attempts at neuroscience inspired AI. Drawing these connections also enabled us to offer a number of suggestions for future theorising in the neuroscience of memory.

Chapter 3 — The brain as twofold object: a rough theory of biological computation with consequences for representation and memory traces

Abstract

It has been assumed but not proven that neural activity is the vehicle bearing representational content. I argue this stems from the way existing notions of computation are contorted to apply to biology, and put forth an original alternative interpretation of neural computation that takes real biology into account, specifically the causal topology of basic cell-signalling mechanisms and the lack of partitioning between memory and processing in real neural networks. I then turn to showing how this view of biological computation complicates existing attempts to locate representations in neural activity patterns, and employ a case study of recent optogenetic causal intervention studies of memory behaviour to illustrate this. Consequences that fall out of this argument include that some existing conceptions of neural representation and memory traces should be dispensed with.

Section 1.1 — Introduction: debate about representation has focused only on neural activity

Naturalising the concept of mental representation using neuroscience is a much contested and reviled project in contemporary philosophy of mind and cognitive science. When attempted, the problem space is typically carved into two main questions—the *representational status question* asks what it is for something to count as a representation against a causal background, and the *content-determination question* asks how the content of a given representation becomes fixed, such that it is a representation of X and not of Y (Schulte & Neander, 2002/2022; Lee, 2019). Central to addressing each of these problems is the issue of "grounding" representations. Representations are typically cast in terms of "vehicles" that bear representational contents by virtue of their causal properties, with a central tenet of representational theories of mind being that representations are individuable physical particulars. The challenge then becomes one of looking to neuroscience for putative candidates that may function as representational vehicles, a project one considered impossible (Fodor, 1974).

Shea (2018) is an exemplar account attempting this within the teleosemantic tradition. He employs examples from the neuroscience of probabilistic decision-making, motor control, place cells of the hippocampus, and recently clusters arising from projecting high-dimensional recordings of neural activity into lower-dimensional spaces (Shea, 2024) in his attempts to ground representations in neural variables. Although he does not argue that representations are only or must necessarily be instantiated in neural activity, "almost all the work that…might vindicate some version of the neural representation hypothesis is looking at firing rates across assemblies of neurons" (ibid).[32] Although the aspect of the

[32] Since it will play an important role in my argument, it is worth clarifying now what "neural activity" means. Neural activity is a loose term that can refer in the most direct sense to the alteration of voltage across the membranes of neurons and their transient "spikes" as well as to is various derivates collected through electrophysiological probes to produce numerical arrays that can be turned into raster plots or be transformed and visualised for human interpretation by dimensionality reduction mappings. It can also refer to less direct recordings such as ECoG and EEG, and proxies of neural activity such as BOLD activity and their derivates such as data processed through Representational Similarity Analysis. It is an interesting historical issue as to why "neural activity" came to have exclusively this meaning and not include all of the ongoing molecular/cellular/genetic activity of the brain.

representations debate I wish to focus on here is the issue of grounding the vehicles of representations in biology, rather than debating the use-criteria[33] of representational content, it is worth noting that disagreement over when criteria have or have not been met by different accounts also involves discussion of neuroscientific details that are based off data derived from or models of neural activity. Krakauer (2021) argues that one candidate of a representational vehicle used by Shea—that of a cognitive map instantiated in place cell firing properties of the hippocampus—is obviated by the discovery of the successor representation algorithm (Mommenejad *et al*., 2017) in computational neuroscience which accounts for learning but does not invoke a model used by the system. Another example, Burnston (2021) argues that neural vehicles cannot be individuated as envisioned by Shea (2018), citing the example of how the analysis methods in macaque electrophysiology (e.g. Principal Component Analysis and Linear Discriminant Analysis) leads to population constructs that are highly distributed, meaning that "one cannot ascribe contents corresponding to a given task dimensions to any single cell or group thereof." In other words, criticism of existing frameworks in refining the literature on representation involves engagement with areas of experimental neuroscience literature that analyse neural activity, either reinterpreting the studies and literature cited, or reinterpreting the results of the data analysis methods used to discover patterns in neural activity.

Important for discussion here is that there is more to neuroscience than the study of neural activity. This includes fine-scale neuroanatomical characterisation of the brain and molecular and cellular neuroscientific methods (used at any scale). Not all of neurophysiology is electrophysiology. What else in neuroscience could be relevant to representation? Criteria for what it takes to count as a representation are contested, but in general stipulate that representations *stand in* for something else in an *uncoupled* manner, that is a standing-in even when that which is being stood in for in online sensorimotor engagement is absent (Orlandi, 2021). Given the centrality of the uncoupling condition for what it takes for something to count

[33] Any attempt to explain representation must meet the "job description challenge" (Ramsey, 2007) —for something to count as a representation is must be *used as* a representation. There are other issues, about whether representations can be non-conscious/subpersonal, nonconceptual, and doxastic.

as a representation, it is surprising that discussions of memory and memory traces have not featured in the discussion. One could give a gloss to the effect that the concept of representations and memory traces share a core commonality as memory traces "stand-in" for previously learned semantic associations or complex sensory experiences no longer present, thereby meeting the uncoupling condition. Memory traces and representations are not just conceptual facility members, as representations do not spring forth from aether. Where are all the representational vehicles that are not presently being used? If a representation can be "called up" (to use computer science language), from whence does it come? Presumably their directly necessary conditions persist in the brain, however the literature on neural representation does not appear to enter into dialogue with the contemporary neuroscience of memory or the philosophy of memory. Shea (2018) does not index 'memory' in its glossary, while Neander (2017) provides only cursory acknowledgement of memory traces.

Not only does debate over present attempts to ground representational vehicles amount to a discussion about whether neural activity patterns do or do not meet important philosophical criteria, but no argument is offered as to why attempts to understand how neural states have meaning or carry content should be limited to neural activity and its human-generated data-object proxies. In short, no other kind of neural data is given this level of significant treatment, and I know of no argument for *not* drawing from other kinds of experimental neuroscience data. I take this as sufficient to argue that there is an implicit assumption in the philosophical literature that neural activity is where representations will be grounded. To wit, that drawing from studies of the transient physiology of firing neurons is what will help develop and settle these issues.

There are two distinct objectives running through this chapter. Despite framing the preceding discussion in terms of representation and neural activity, the first objective is the articulation of a general view of biological computation. Unlike digital or analog computation, I seek to establish a bespoke account of computation that places primacy on attending to the unique physicality and causality of biological processes over borrowing heuristics from theoretical

computer science, mathematics, or engineering. This comprises the longest part, Section 2.

The second objective is to argue that the focus on neural activity is limiting. I do this using a combination of arguing from a view of its incompatible with biological computation, as well as from a case study from contemporary memory systems neuroscience. More specifically, I use the case study to show that those features of nervous systems appealed to to naturalise or "ground" the concept of representation are definitively *not limited to neural activity* but can include neuroanatomical properties such as network architectures that contribute to the determination of online mental content. This is Section 3.

These two objectives go hand in hand. While introducing a theory of biological computation is not a necessary step in order to argue for the importance of looking past neural activity, I spend significant time articulating the view of biological computation because, among other purposes, it rapidly contextualises why and how the case study I have chosen *deflates the explanatory privilege of neural activity* in the search for representations and constrains the scope of philosophy of neuroscience and mind in general. Thus, long-standing issues about representation and memory may be regarded as a test-case for my account of biological computation, and although they are presented together, readers may find the view of computation compelling but not be convinced that attempts to ground representations should look beyond neural activity, or may be convinced by the latter but disagree with tenets of the view of computation. Ultimately, the intention behind this structure is to create a theoretical and experimentally informed argument that brings about an additive effect in upending the focus on neural activity. This significantly complicates the philosophical and scientific project attempting to ground representations. However, though I have striven to ensure both my presentation of biological computation and the case study are agnostic as to whether representation must be couched in terms of vehicles and contents, this project is not incompatible with vehicle realism and in fact opens a fertile ground of further philosophical research and collaboration with neuroscience.

Before introducing the theory of biological computation, I attempt to rationalise this near-exclusive focus on activity by linking it with the broader influence of computationalism.

1.2 — Neural Computation is done by activity dynamics, linking RTM within CTM

Why is there this exclusive focus on neural activity in attempts explain representation in neuroscientific terms in the philosophical literature? One answer is that in neuroscientific papers the term "neural representation" is used to refer to patterns of neural activity, so when philosophers read neuroscientific papers or attend talks and encounter the word representation, they may inadvertently co-opt the same substrate in their arguments. For example, patterns or dynamics may be said to "represent" a particular stimulus ("stimulus representation") or a particular decision ("representation of decision variables"), and famously that place cells firing in the hippocampus "represents" location in space. This usage of the term overlaps with aspects of but differs from how representation is variously understood within philosophy. The conception of "neural representation" rife in neuroscience is ultimately a false friend for that of "mental representation," as the former is used typically for any observed neural variable that covaries/correlates loosely with features of an external world. However, to say that attempts to ground mental representations have focused on neural activity because neuroscientists uncritically refer to patterns of neural activity as neural representations is not a very satisfactory answer to the original question. It simply prompts the question of how neuroscientists came to use the term representation in this way.

A second possible explanation for the source of the focus on neural activity is that it comes from the influence of computationalism in both philosophy and neuroscience. Representational theories of mind (RTM) overlap strongly with

computational theory of mind (CTM). Representation may be seen as central to computation (Fodor, 1981), although computation can be defined without representation (Piccinini, 2008). The history of the origin of computers and so-called artificial neural networks saw attempts to establish connections with neuroscience. The comparison between Turing's computer and the brain was made in McCulloch and Pitts (1943) while von Neumann's (1945) Report on the EDVAC used the term "memory" when referring to a unit storing symbolic states in the first presentation of the "von Neumann" architecture. Attempts to equate the brain with a computer and tease out possible similarities was a central project and legacy of McCulloch's career (Kay, 2001; Piccinini, 2004) as well as von Neumann's (1958). That brains compute is foundational maxim of contemporary neuroscience (Marr, 1982; Koch, 1999).

Computationalism or the brain-as-computer metaphor has been and remains much criticised in both philosophy and neuroscience (Dreyfus, 1965; Varela, Thompson & Rosch 1991), with consideration of alternatives remaining an engine of fruitful new ideas (e.g. Kelty-Stephen *et al*., 2022). Nonetheless, the language of computation remains the worldview neuroscience papers are couched in. This covers any casting of brains or nervous systems as "information-processing" systems.[34] Non-computational neuroscience papers commonly define their objectives and contextualise their findings in terms of furthering understanding of specific "neural computations." With the growing depth of computational neuroscience theory and influence of machine learning on neuroscience, I see "neural computation" as set to continue as dominant scientific parlance for the foreseeable future. In this context, "neural activity" is interpreted in terms of the computations it is potentially performing or capable of performing. A gloss of arguably the most prominent contemporary perspective is that neural computation is computation over low-dimensional manifolds of population-level dynamics of neural activity (Ebitz & Hayden, 2021; Barak & Krakauer, 2021; Khona & Fiete, 2023). Although the following associations are not linked through philosophical

[34] Piccinini and Scarantino (2010) hypothesise that the conflation between "computation" and "information processing" originated in the usage of both terms in the early cybernetics literature. This paper is principally concerned with computation.

rigour, one could uncontroversially claim within a contemporary neuroscientific context that since neural computation is assumed to be done by the dynamics of neural activity, and use of representations is a class of computation, *representing* is going to be a particular class of neural activity dynamics such that some dynamics can represent and others do not.[35] For example, in mice computing the head-direction angle is done through a low-dimensional ring topology in neural activity (Chaudhuri *et al*., 2019; Langdon Genkin & Engel, 2023). The head-direction angle could be deemed the representational content and activity-state the vehicle. Thus there is an interesting tension whereby the brain is not obviously like a classical digital computer but the language of computation is entrenched in the study the brain. What then is computation? And when neural computation has been argued to be different (e.g. Piccinini & Bahar, 2013), how exactly is it different? I first attempt to clarify different examples of what computation in general can connote, before turning to why they are inappropriate to apply in biological contexts, including neuroscience.

Many other systems besides brains are described as being capable of computation or performing computations—computers, organisms, artificial neural networks, neuromorphic chips, quantum systems, and societies—often despite the lack of a clear definition of what is meant by "computation." Computation may be defined as problem-solving (the "computational problem"), as the application of a function to an input to produce an output (the "model of computation"), by comparing computers to calculators, although ultimately the most common approach to defining computation is appealing to the different gradations of formal models studied in theoretical computer science. Here, the simplest computing system is any system equivalent to a look-up table (LUTs), followed by Finite State Machines (FSMs), with the most general definition being any system capable of Universal Computation as outlined by Turing (1936), with the most powerful computer being anything equivalent to a Turing machine, such as a von-Neumann architecture computer.

[35] This can be seen in a random selection of recent neuroscience articles (Bernardi *et al*., 2020; Courellis *et al*., 2024).

However, scientists at the intersection of neuroscience, machine learning and neuromorphic computing have argued that the standard theoretical model of computation—that of Turing-computability and von-Neumann architectures—has outlived its usefulness for understanding neural computation.[36] Citing well known issues about the fact that neural systems are at best analog, do not perform serial processing, are unclocked and their computations are irreproducible (Jaeger, Noheda & van der Wiel, 2023; Aimone & Parekh, 2023), they argue for the need for a new bespoke theory of neural computation. This is what I hope to provide through way of articulating an account of biological computation.

[36] Their approach therefore challenges any curtailing of the semantic range of "computation" to Turing's definition.

Section 2 — An alternative view: Biological Computation

Alternative definitions of computation applicable to the brain have been articulated within philosophy. For example, Piccinini has developed a definition of physical computation as the "manipulation of medium independent vehicles by a functional mechanism in accordance with a rule" (Piccinini 2020), while Maley (2011) provides an account of analog computation applicable to the brain (more on analog computation in a moment). Like Piccinini, I will argue that neural computation is its own kind worth understanding on its own terms, that is I support a view of neural computation that is neither digital nor analog but *sui generis* (Piccinini, 2020). I too also follow the emphasis on situatedness, and the view I will articulate below is also one that does not see computation and enactivism as contradictory or mutually exclusive. Where I will differ in my conception of computation from Piccinini is that I am uncomfortable with a definition that includes anything resembling "medium independence" or "rules" as fundamental. Rather than focusing on neural computation, I develop an account of "biological computation" which I treat as its own kind that does not recourse to computer scientific or mathematical intuitions,[37] and within which neural computation is but a specific limiting case. This alternative and *more biological* view of computation will change undermine our attitude towards neural activity and therefore extant attempts to locate representation vehicles therein.

2.1 — Scales of biological computation

First, I start with a general framing of the broadest gradations of computational complexity for biological matter. Provisionally, let the space of biological

[37] While this may seem like cutting the final tethers needed to justify using the term computation at all, I will argue otherwise. There are several things the following theory of biological computation is not. First, it is not part of an attempt to view everything as computation, such as the pancomputationalism advocated by Wolfram (2002). Second, while many of the following principles defining biological computation I will discuss are issues partially or similarly discussed in areas such as artificial life, in particular the simulatability or computability of life (Chemero and Turvey, 2008; Varela, Maturana & Uribe 1974), computational modelling or the possibility of simulating of life is not what I will mean by "biological computation" (Piccinini, 2007). Thirdly, this theory of biological computation has nothing to do with computational biology *à la* the various -omics, which are a form of applied data science.

computation refer to the number of *possible configurations* that biological matter can take on, defined with respect to arbitrarily chosen "levels of organisation."[38] Second, I assume none of the previous definitions of computation (problem-solving, input-output, universal computation etc.). I break biological computation down into three broad classes in descending order of computational power:

i) evolution

ii) development

iii) learning (e.g. neural plasticity)

Here, we are differentiating the computational power of biological matter on the basis of timescale rather than architecture *per se*, although it is also more than just timescale. In terms of picking the boundaries, there are two tiers of constraints that form the bounding criteria for drawing the distinctions in this manner, these are the "barrier of individuality" and the "barrier of mitosis" (Figure 1).[39]

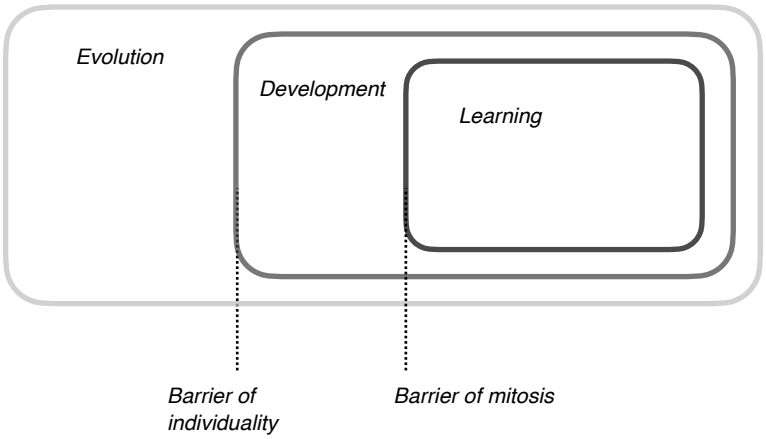


Figure 1: Learning/neural plasticity is doubly constrained due as cells are bounded within an individual and neural tissue is post-mitotic. Development sits in the middle because, although there is still mitosis, cells are bounded within an individual. Evolution is the most powerful because it is neither bounded by limitations to cell division (for the sake of within-lifespan memory) nor to what is possible during the single lifespan of any individual.

[38] I say provisionally, because as argued by Kauffman (2019), the number of possible configurations of biological matter is "unprestateable." It is not possible.

[39] There are several issues with this overly simplistic subdivision that are important, although not for the purposes of this article. First is that not all learning is neural plasticity, and the second is that there are many different conceptions of individuality in theoretical biology (e.g. Krakauer *et al*., 2020). In addition, it should be noted that this is not meant to be a direct comparison to the three categories to Turing Machines, Finite State Machines, and Lookup Tables in theoretical computer science.

The remainder of this section follows the following general order. First, I outline the core principles of biological computation *in the negative*—through demonstrating how it differs from classical conceptions of computation. I then offer a positive presentation. Throughout I will focus on how, with this general reference framework in mind, each of the classes follow the same general principles regardless of the scaling in computational power as constraints are lifted or varied. The result of this is that while we see learning to be the weakest—a *limiting case* out of the three classes of biological computation—it is still follows the principles outlined, making applying a classical view of computation to the brain unwarranted. This will anticipate the possible counterargument that neural computation can be "locally approximated" by classical computation in a meaningful way.

2.2 — Approaching biological computation in the negative: undoing two cuts made at the core of computer science and mathematics

I leverage two primary points in parallel as prongs to open the problem in the desired way. These go together like the sides of a coin and are not independent. The two issues I wish to draw attention to is the fact that in biology there is

i) No symbolic abstraction
ii) No physical partitioning between memory and processing

2.2.i) The first issue issue: partitioning between symbol and structures that operate on symbols

What could be said to be the criteria for something to be symbolic, to *function symbolically*? In the context of classical digital computation, one definition is the

symbols "code for" variables.[40] Digital computers are also symbolic from the ground up and in a way that, like Jaeger (2021), I locate in a lineage descending from a powerful separation introduced in Aristotelian syllogistic logic. Specifically, I refer to what could be called *the separation between the form and content of an argument*, such that due to appropriately following a series of defined rules, "correct" inferences can be arrived at from initial premises, *regardless of what the argument was about*. One can argue syllogistically for awful things and still have a "correct argument" within this framework (soundness/validity distinction). Today, even the rules/algorithms themselves can be specified and executed on a variety of devices, not by magic but because of a shared cultural milieu and agreed conventions between system designers and manufacturers. The result of all this collective effort is that computation in digital computers is able to proceed almost completely dissociated from nature. The systems are designed not to care because they do not have to—their "needs" are taken care of by the human design and manufacturing labour that brings them into existence and the biosphere-supported-electricity grids that (at least at present) continue to supply their power. Life, in contrast, does not have such a fundamental partitioning at its origin.[41] A basic education in biology involves covering the processes by which organisms support themselves through the stripping of electrons from donor compounds,[42] whereas in computer science "Students…must do coursework in formal logic, not physiology; and their theory textbooks speak a lot about logical inference steps, but never mention seconds" (Jaeger, Noheda and

---

[40] In digital computers, the magnitudes IIIIIIIII and IIIIIIIIIIIII become "coded for" by symbolic binary strings, but 1010 is not physically "smaller" than 1111. For an example of this separation in mathematics, it is not as if the number 5 on the dice is intrinsically especially heavier, as this would violate the basic properties defining the natural numbers.

[41] The fundamental partitioning in life in the form of cell membranes is different and exists for the reasons of physics, not symbolic operations: for creating the energy differentials and chemical separation necessary for particular reactions. There was never the creation of abstract symbols and rules that operate on them. Not even in the turning of RNA into proteins by ribosomes counts, since RNA strands can take on various secondary structures that have unique functions. Even the "symbolic sequences" can interact with other molecular processes because they are always physical.

[42] In theoretical biology, following Maturana and Varela's (1972/1991) conceptions of autopoesis and closure, Moreno & Mossio (2015) and Kauffman (2019) have argued that most of the "work" (for which we could also substitute "processing" or "operations") of life is creating the (internal and external) conditions for life to derive energy to allow them to create the conditions to derive energy, and so forth.

van der Wiel, 2023). Living matter never separated from the natural world. Biological computation and surviving are not dissociable.

The other side of symbolic abstraction in digital computation is that symbols are operated on by something other than themselves. Turing envisaged the computer as a person, a mathematician "writing certain symbols on paper" (Turing, 1936). The feat of engineering that gave us computing as we know it today was to externalise not just the symbols but the manipulation of these symbols in accordance with a highly particular style of explicit reasoning into an apparatus capable of operating on those symbols.[43] Regardless of whether the "processor" is understood as a human operator or machine, Turing's original imagery is illustrative of the separation between symbols and that which operates on them. I will refer views of computation that adhere to this fundamental split as "classical computation."

2.2.ii) the second issue: partitioning between memory and processing

What else could be said to be the criteria for something to function symbolically? Symbols, regardless of what they "code for", *once written must stay the same*. In Turing's original image, this is already accounted for by the use of a preservational medium such as physical paper to write symbols on. This was mechanised by the development of materials (e.g. transistors and magnetic tape) which work by effectively bringing time within them to a standstill, that is by "defying the mutations of time" (Jaeger, 2021). The purpose of preventing corruption of "stored" symbolic states is to carry the results of certain past computations into the future. This leads us into the second core issue, that of memory.

In digital computers, not only is there a dissociation between symbols and the structures that process them, but also between "memory" and "processing." The

[43] The human was not eliminated, rather our role, even if not a computer engineer, became the setting of initial conditions and interpretation of outputs.

site of memory differs from the site of computation. Memory here bottoms out symbolically as binary states held in inert addresses (and in some cases in an external repository), while all computation is done by the CPU. In von Neumann's (1958) words "forms of memory…are technologically entirely different from the basis active organs of the machine…" This separation leads to the following seemingly innocuous question, where are memories to be found in biology?

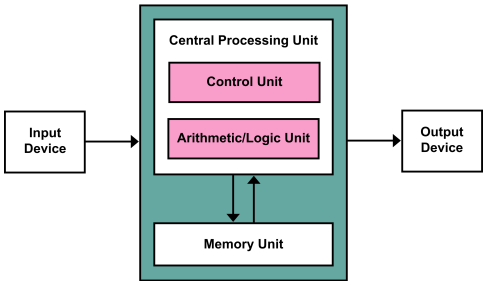


Figure 2: The von-Neumann architecture.

> "We know the basic active organs of the nervous system (the nerve cells). There is every reason to believe that a very large-capacity memory is associated with this system. We do most emphatically *not* know what type of physical entities are the basic components for the memory in question." [italics in original]
>
> — von Neumann (1958, p.67), *The Computer and the Brain*

This framing of memory being separate from computation predates the origin of computers, e.g. the boons of storing the results of complex calculations so they did not have to be performed again can be seen for example in the development and distribution of logarithmic tables, while the idea of a "memory trace" has been a recurrent fixture of western philosophical and now scientific speculation (for reviews see Robins, 2023 and Najenson, 2021).

An important aside is that connectionist networks use a fundamentally different architecture to a von Neumann computer. Removing the separation (or most of it) between memory and computation (to give "in-memory computing") has been part of the development of connectionist models and is now a physical reality in

neuromorphic hardware, so do they count as performing biological computation? Dispensing with the idea of "memories" understood as abstract nouns located at an address or as imprints in a material whose purpose is to record and only record for a separate reader is certainly a conceptual revolution hard to grasp and one neuroscience is arguably yet to fully reckon with.[44] For example, there is still the belief that learning leaves distinct traces corresponding to the encoding of new information, and that these traces serve as records of what is learned and can later be read out (Gallistel & King, 2010). However, identifying such literal candidate traces within biology has so far proven unsuccessful (see Chapter 1). I will return to this point towards the end of this piece. Despite claims that simulating ANNs on classical digital computers has no bearing on their ability to perform non-classical (and therefore potentially more neural-like) computation (y Arcas, 2023), ANNs still do not overcome an initial abstracted partitioning from the world mentioned above. There is still a primary bracketing-off from natural processes, because all inputs artificial networks take in must be converted/assigned to abstract vector embeddings. Living organisms, in contrast, are always directly coupled with the world and it would not make sense to talk of or look for "embeddings" in biology.[45] Ultimately, ANNs solve issue ii) but not i), therefore they do not do biological computation. The solution to i) and ii) in biology is the same, and this holds true for single cells and entire populations as well as neural tissue.

2.3 — Biological computation articulated in the positive: Biological computation is DRG (Displaced, Reflexive, Groundless)

In contrast to classical computation, how should we describe biological computation? Informally

---

[44] Indeed, it may not be appropriate to use the concept memory in the singular e.g. "a memory" in the context of ANNs and biological networks, because it connotes an abstract object sufficiently clearly bounded to be referred to as separate from the "substrate" it is "implemented" in.

[45] This may be the starting place of the "barrier of meaning" in artificial neural networks (Rota, 1985), but this is not within the scope of this piece.

A biological system does not perform operations on symbols,
it performs them on itself through displacement,
that is it performs operations on itself through a spatiotemporal intermediary.
There is no simple duality between fixed structure and symbol.
The past is preserved not by encoding results and writing them to memory,
but by the organism altering "the ground on which it stands."
An organism does not "store" its past anywhere, it is its past.
It does not compute predictions about the future, it creates future.

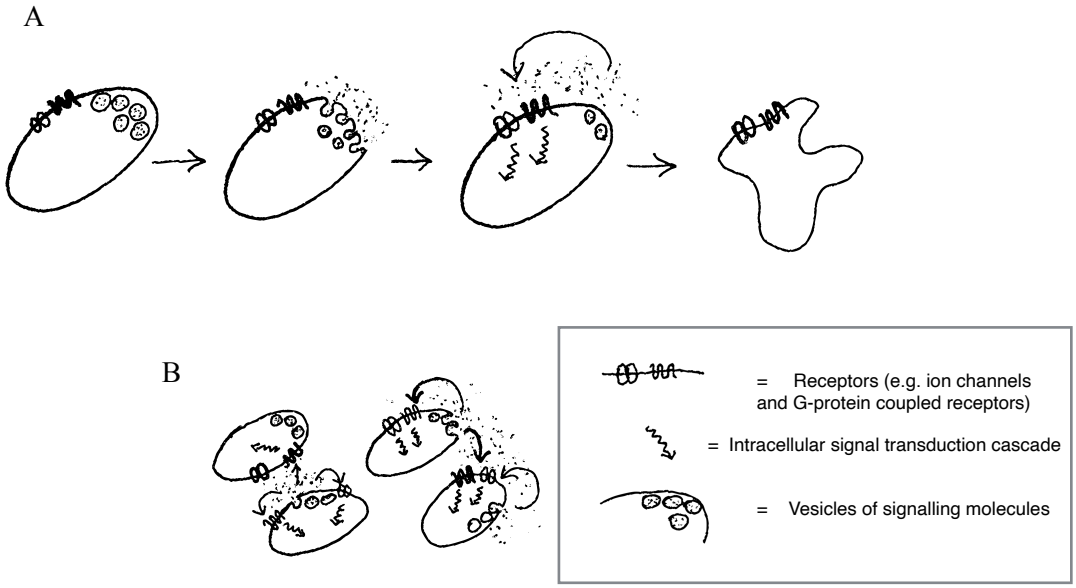
2.3.i) — Biological computation is reflexive: operation is self-operation, not input-output operation

A biological system does not perform operations on symbols, it performs them on itself…

As alluded to, I will argue that life does not operate by employing a kind of primordial partition or abstraction between symbols and structures that operate on them in accordance with defined rules. Computers and artificial neural networks do not operate on themselves, but on inputs from something these theories call "the world," and return to "the world" things or states called "outputs" (Hurley, 1998). The connotations of this are not obtuse, it is simply "to take in" and "to put out." The topology of the causal process in biological computation, however, is fundamentally different. For a singe-celled organism, I will use the most straightforward counterexample: *autocrine signalling*. This is the production and release of signalling molecules by a cell that in turn act on the cell's owns receptors to effect a particular change.[46]

[46] Other intracellular signalling mechanisms can be used, it is not necessary for molecules to cross the membrane and then cause signal transduction cascades count as operating on oneself, rather it is the most clear illustration.

Figure 3 — Autocrine Signalling



This means not that there is no difference between input and output, but that terms like input and output simply do not apply. Biological systems operate on themselves. This is *reflexivity*.[47] In contrast, western systems of logic tend to abhor (nay, do not get to work in) the presence of (apparent) contradictions of so-called "self reference." For example, without including additional axioms to naive set theory to address Russell's paradox—that *a set cannot be a member of itself*—there is no first order logic and no foundation for the majority of mathematics used today.[48] Biological systems in contrast *do not care about this*. Indeed, living matter has been highly successful, colonising and transforming the planet, all while flouting the rules of set theory. What is *verboten* in western logic is not a problem in the living world. This issue may stem from the fact that the ability to operate on symbols requires the specification of hard rules for what is possible (the types of operations allowed), and that while axiomatisation makes possible various architectures capable of computation, it closes off certain other possibilities, such as computing through the transformation of matter. Attempts to circumvent this, such as applications of the cybernetic insight of "feedback" has

[47] Reflexive here and throughout does not mean self-generating like the term autopoeitic.

[48] I apologise to logicians and mathematicians for such a fast and loose usage of an example of paradoxes associated with self-reference.

brought new techniques (Recursion, Markov Processes, Autoencoders), but the fundamental hurdle cannot be overcome. The issue is that when passing the outputs back in as inputs, when it comes to material instantiation these effects are typically limited to the level of the simulation, the hardware performing the computation is not changed by it. However, biological matter does change when it operates on itself, and does so irreparably.[49] There is surrender and sacrifice to a process of change in biological computation. To summarise this point, if you try to operate on yourself but remain confined to physical parts with exceedingly few degrees of freedom compared to organic matter, there are certain limits. It is never possible *in toto*. However, if you are not limited to the symbolic realm, to the operation on symbols by stable structures, there are no issues with operating on yourself. You are both the operator and operand, or rather neither, because this framing does not apply (I will return to this in a moment).

This is what separates my account of biological computation from that of analog computation. Maley (2021) has argued that analog computation provides a more appropriate sense of computation for understanding the brain because it takes advantage of, rather than abstracts away from, the physicality of the mechanisms used. I concur, but biological computation must go one step further than analog computation. There are huge design and materials engineering challenges to get analog computers to perform complex operations (e.g. difference engines are very large), and they cannot alter the material they are made of, regardless of whether these are a purely mechanical or electronic analog computer, just like digital computers. Biological systems do not have these same constraints. While it may be possible to consider the brain as approximately analog over short timetables, whereby the analog mechanisms are defined up to some threshold of turnover to be an invariant of unfolding material changes, I would resist this as it requires ignoring the fully reflexive groundlessness of biological computation that differentiates it from digital and analog computers. Taking "species" to be the

[49] Of course, there is no stopping someone from redescribing autocrine signalling in terms of inputs and outputs. The state at t_1 could be described as an output that is the input to the state at t_2. However, I take this to be a framing that overcomplicates the biological, and adding extra conceptual baggage by trying to describe the process like a Markov chain. Biological computation is more parsimonious.

most general biological kinds, we see the depth of Darwin's insight that *species are not immutable*.

Let us return to the example of autocrine signalling in Figure 3. The result of this displaced operation on self is change to the cell performing the operation. Perhaps it will never perform the operation again, or it will perform it differently from then onwards or for a period before reverting. The first consequence is that *there are no fixed units of processing in biology*. Not even the genome, often and erroneously compared to a book, is stable within an individual lifespan (whatever about within evolutionary time), even within the 'simplest' of organisms. A single bacterial genome for example extensively and radically shuffles and reshuffles itself (e.g. Wilbanks *et al*., 2022).[50] Even highly conserved proteins do not count as physically stable processing structures, since they are constantly being post-translationally modified. Thus, even if we limit the scope of biological computation to a single cell, without replication, it can still be understood as operating on itself, as altering "the ground on which it stands." This to me is the fullest interpretation of Darwin's *descent with modification*. In biological computation, there is no solid ground, *nowhere to hang our hat*. This is what is meant by calling biological computation *groundless*. This becomes even more apparent when we include replication to look beyond individual lifespan and examine populations on evolutionary time—speciation and extinction unfold.

The final characterise of biological computation I wish to highlight is *displacement.* This explains why self-operation is not solipsistic. While the brain can be described as communicating with itself using self-signalling, and this self-signalling is sufficiently complex to also coordinate motor movements while doing so, the use of an intermediary allows sensitivity to other influences through coupling of the intermediary with other forces. This principle can also be illustrated through autocrine signalling. Because a cell signalling to itself through

[50] The circuitry of the ALU or DRAM is not excised, chopped and changed while executing a given program to enable it to run new types of operations in the future. We humans do this to the circuits on the level of decades of technological evolution, and is thus a mute point. In analog computers, the cogs of the differential analyser were not melted down and reformed on the fly in order to perform different calculations.

the exocytosis of signalling molecules into the surrounding medium does not preclude another cell also responding to these molecules (if competent), autocrine signalling can also be paracrine signalling. A population of cells each signalling to themselves but also being able to respond to compounds from others is what enables *quorum sensing*, a fundamental "computation" of bacterial populations, to work.[51] In other words, externalising the intermediary of the self-computation process allows self-operation to interface with—to be partially stitched into—what is beyond any individual cell. In the case of the brain, neural activity allows the brain to interface with the body. We will return this later in the section on redundant descriptions, because we can also describe worlds as operating on brains through exerting influencing on neural activity.

In neural tissue, how to understand self-operation and *no fixed units of processing* can be understood as follows. While it is possible to imagine a particular "brain structure" or "neural circuit" as the fixed "processing structure" and describe it as performing operations on arguments/operands, implementing so-called "canonical computations" (Carandini & Heeger, 2012; Miller 2016), these are also not like classical computations.[52] Certainly, neural circuits can be described as passing sequences of spikes along and transforming them, but importantly, *these activity patterns and sequences can induce changes in the circuits themselves*. This changes the very structure that previously "processed" them, leading to different results in future.[53] In other words, a "processing structure," such as a subsection of the connectome, may be described as a circuit with unique architecture C1 that performs a certain unique "operation" O1 such that after C1—O1 we now have C1' and later O1'. Suddenly we have C1''''''''''''—O1'''''''''''' after which they might as well be called C2 and O2, having at some arbitrary point undergone a

---

[51] An example being the phase transition from isolated into a colony.

[52] The neuron doctrine here is especially unhelpful. The concept of a "functional unit" can imply looking for a "processing structure" that is invariant. Rather, we should be wary of assuming biological "function" implies something not allowed to vary in its nature.

[53] The only computer science version of this may be overclocking or overheating a CPU and oxidising transistors, impacting future computations by damaging hardware. The biological comparisons may be looking directly at the sun or oxidative stress, which are not relevant to the discussion here.

transformation in kind.[54] I integrate these insights into the following maxim: *the brain (or a given circuit) does not perform operations on symbols, it performs operations on itself through mediators, between regions and versions of itself in time*. It does this operation on itself primarily *through* electrophysiological activity as the mediator or intermediary, but also glial processes and neuromodulator release as other intermediaries. In other words, neural activity here is not granted any privileged status, we see it as intermediary in an unfolding computational process.[55] Where previously it was treated as performing operations "on the world" it is now seen as one way the brain operates on itself. It does not take much to draw out how this accounts for there being no partitioning between memory and processing. Trying to approximate neural computation as classical computation requires ignoring this principle of self-operation.

2.3.ii) Generalised descent with modification encapsulates "memory"

Consider again C1 and O1. We see that "structures" do things, and that doing alters the structures under select conditions such that they may do differently, and not just "any differently," but in a way that may be potentially useful. I present a second maxim derived from the insight of self-operation. *The brain too takes the past and brings it into the future not by encoding and writing it to a memory address and reading it later, but by altering its own physical architecture, the ground on which it stands*. There is no difference between memory and processing structures, the past makes processing possible. In short, just as there is no need to apply the conceptual framing of a brain operating on "inputs" from something called "the world," nor is there a separate structure created within the brain that is used to store the results of past operations. Just as there is no evidence that spike are symbolic, there is no evidence that the brain carries the results of past

[54] This is not to say that limitless variation is possible within an individual lifespan (Rule & O'Leary, 2022).

[55] There is no evidence that spike trains are symbolic codes, action potentials are not *bits* (Brette, 2019). Just because information theoretic analyses can be performed on time series data, and neural activity can be recorded as time-series data, does not mean that spike trains carry Shannon information, regardless of whether spikes are understood as bits (as initiated McCulloch and Pitts, 1943) or otherwise.

computations into the future by writing to a structure specialised for storage, or that what is written in is symbolic states (Gallistel & King, 2010). No abstract code has been found in activity, nor has an abstract code been found in the connectome.

While genetic mechanisms could also be suggested as a source of stability through time, the timescales of interacting with the genome are not sufficiently fast for the coordinating the movement of bodies in space. Evolution likely achieved this—a level of organisation above dynamically stable genomes—through halting normal cell division and rendering neural tissue post-mitotic (Figure 1), allowing connectomic structures to "defy the mutations of time" for the duration of an organism's lifecycle.[56] The connectome makes possible and stabilises brain dynamics, while the connectome, previously highly dynamic during development, is made possible and stabilised by the genome (or rather, the inhibition of developmental signalling pathways). The connectome is a process too, just one that where the dynamics become more restricted after development. However, in cases for organisms that undergo metamorphosis, the connectome can be reconfigured far more dynamically than can be done with learning by a second phase of development.[57]

There are multiple mechanisms for descent with modification, for carrying forth the past, and what specific mechanisms are purposed is likely determined by ecological timescales. Biological matter can compute by moving through space and as well as morphological transformation (e.g. ant colony navigation, slime moulds, the immune system, neuronal axons migrating during developmental to ensure viable wiring of the brain). But biological tissue whose movement and morphological transformations are limited can also compute through releasing signalling molecules into a fluid to neighbours, shared cytoplasm via gap

[56] However, this is not the same as bringing internal time to halt as is done in computer memory materials. The biological equivalent of a computer-like memory system that brings internal time to a halt would be the freezing and thawing of million year old fungi (D'Elia *et al*., 2009) or recovering a tardigrade from the vacuum of space (Jönsson *et al*., 2007).

[57] Given that the development process already involves transitions of previously evolved abilities from genetic mechanisms into neural tissue, it would not be surprising for learned dispositions to be carried through morphogenesis through genetic mechanisms and reinstated in remodelled neural tissue.

junctions such that variation in chemical concentrations passes through coupled cells (e.g. astrocytes), and by varying membrane potentials across a network (e.g. neurons, myocytes). Importantly, biological matter can transition from computing morphologically to one that subdivides the tasks of movement and control. At no point does biological computation de-cohere, renormalise, or phase transition into classical computation.

2.3.iii) Redundancy of multiple operator-operand constructions in biology: "The brain as twofold object"

There is a lingering issue with trying to state that there is no fixed distinction between structures and what they operate on in biology. A seemingly innocuous question that may emerge from this is the following: "is the connectome operating on activity, or is activity operating on the connectome?" Both descriptions seem equally adequate. This apparent equality of mirrored descriptions is also seen at the largest scales of biological computation in *niche construction*. In the general cognitive sciences, "action" usually refers to individuals "operating on the world." In contrast, in neo-Darwinian theory, although evolution is not an entity granted agency, the "action" (albeit passive) is of selective ecological pressures "operating" on populations. The tension that arises in attempts to merge these two different framings has generated many words, but the essential insight of niche construction is that populations evolve in response to environmental conditions, *and* populations are also capable of transforming environments in a way that in turn effects the selective evolutionary pressures acting on the population (Olding-Smee & Laland, 2000).[58]

Returning to the brain, recall the point of using displaced intermediaries in self-operation, that "neural activity allows the brain to interface with the body." Covering more expansive loops, one could just as easily describe the brain as

[58] This makes modelling biological computation with existing mathematical tools such as coupled differential equations difficult as *organisms construct their own boundary conditions* (Kauffman, 2019).

operating on itself through the intermediary of the body or the world. In addition, bodily tissues too operate on themselves, and do not need to interface with the nervous system to do so, they can use local cell signalling mechanisms as well as the immune system. The point that the brain is not the only biological organ that operates on itself is crucial. The body is made of cells and these are capable of operating on themselves. Additionally, along with asking "is the connectome operating on activity, or is activity operating on the connectome?" one could also choose to describe activity as operating on itself, with the connectome as intermediary. This framing might appear appealing for theories that locate all stored knowledge in the connectome, or wiring diagram, but which by itself is seen as inert. However, as I have sought to illustrate, "the connectome" and "neural activity" are not natural kinds or separate entities.

In both these descriptions of organisms and environments, connectomes and neural activity, we find the old dichotomy of matter and pattern.[59] In neuroscience, we speak of "neural activity"—often a contraction of "neural activity patterns"—contrasted against "anatomical structures". To risk being pedantic, I hope to have demonstrated that upon close inspection, when one tries to locate which is the processor and which the symbol/data-structure that is "processed", operator and operand, these constructions cannot be located anywhere.[60] There is no stable ground in living systems that supports such a clean categorisation. To add insult to injury, all of this could have been phrased as the world operating on itself through the intermediary of bodies, neural activity, and connectomes. This descriptive redundancy may seem frustrating, but rather than selecting one side (e.g. "connectome operates on activity" or "activity operates on connectome"), or accepting equivalence of descriptions (both pragmatic strategies), I take this as indication that operator-operand constructions are simply not the most appropriate framing to cast biological computation through. Instead, we see that *biological computation is coreless*.

---

[59] The etymology of matter and pattern is *mater* and *pater*, mother and father.

[60] I suspect that Pattee's (1987/2012) concept of "symbol-matter duality" in biology may have been in part an attempt to deal with this issue. His solution was to propose a duality, that biological matter leads a "double-life" as physical entity and symbol. While this certainly remains a viable approach, the approach here is to do away with preserving monolithic concepts of matter and symbol altogether. At least it is worth attempt to conceive of an intellectual world that is otherwise.

With these basic principles of biological computation elaborated, we can see how this perspective allows straightforward redescription of any attempt to frame computation in biological systems as probabilistic rather than logical inference, including any description of life or brains as probabilistic inference engines using the formalisms of Bayesian statistics (e.g. Dayan *et al.* 1995; Friston, 2010; Clark, 2013). Recall the following statement from earlier: "circuits pass sequences of spikes along, and transform them, but in important cases the activity patterns and sequences themselves induce changes in the circuits. This changes the very structure that previously processed them, leading to different results in future." Recalling the core principle of biological computation of no separation between symbols and structures that process them, there is a way to re-cast inference processes as self-operation: the prior and posterior are not computed by a circuit, *they are the evolving circuit*. The prior and posterior are not distributions "in" populations of neural activity. Rather the system at t_1 is the premise and t_2 the conclusion, and so forth. Thus there can be inference without symbolic abstraction. Biological matter does not "calculate distributions." It is not clear who is "operating" on who, who is the symbol and who is the structure. The material process unfolding is the processing. The matter evolving is itself the computation.

2.4 — But what "is" biological computation?

While computability theory gives one definition of what computation "is" but cannot tell us what to compute with (people, boxes, transistors), up until now I have not given a definition of what this form computation "is" but have only focused only on characteristic properties at three scales of complexity—that it is *displaced, reflexive,* and *groundless*—by paying attention to the unique physicality and causal architecture of living systems. It offers a thoroughly anti-essentialist ontology. However, I have not typed different forms of computation according to this framework, and more significantly, I have not offered an answer

to the following challenge: what would be an example of a cellular or organismal process that is *not* computing, or is all life computing, all of the time?

One approach is to define life itself in computational terms as "problem solving matter" (Krakauer, *personal communication, August 20, 2023*; Krakauer & Kempes, 2024). This uses the "computation as problem-solving" definition and draws an equivalence with living systems. However, computation as "problem-solving" only gets us halfway. The only *universal* "problem" to be "solved" by life is survival/persistence ("universal" in the sense that it must be solved at all timescales, that is evolutionary, ecological, and moment-to-moment). As long as survival is taken care of, living matter is free, free to explore *possible configurations* which may or may not in the future be purposed "for" or as a piece within a "mechanism" that holistically solves a particular problem. In other words, living matter is only computationally lower-bounded, it is not clearly upper bounded.[61] This makes specifying a problem to be solved for every single biological process observed difficult. Recalling niche construction, organisms alter their environments and can themselves be the environments of other organisms, so simply observing an environment without life does not reveal the problems living matter must solve. *Biological matter creates the problems it has to solve as it computes*.

Defining biological computation as problem-solving also introduces a certain observer-dependent carving, namely that we must conjecture our own standards for when something has been computed over a given interval, as well as stating what the boundaries of the problems are. Deriving examples from "life itself" is difficult.[62] While this reverse engineering of problems life is purportedly solving could be attempted by looking to structures or "mechanisms" that perform "biological functions" (e.g. bacterial flagella motors for propulsion), these

[61] Survival is only boundary condition that I see which can be stated without controversy. Although bioenergetic barriers may also be considered an upper bound on biological computational complexity, these boundaries do not apply universally. In the case of the transition of biological matter from prokaryotic to eukaryotic through endosymbiosis, the previous upper bound was simply broken.

[62] In addition, applying computation-as-problem-solving in neuroscience would potentially have too much in common with Marr's computational level, which is not a framework we want to traffic in on account of it exemplifying classical computation (Maley, 2021).

functions or mechanisms are overlapping, uncountable and constantly being repurposed. Thus, altering the terms of discourse does not resolve the fundamental hurdles to computational description posed by the groundlessness of living processes.

If specifying biological computation in terms of problems was insufficient, defining the operations available is even more difficult. In logic, one must know which operations are allowed to be performed (AND, OR, etc). While the "rules" of physics and chemistry could be interpreted as defining a list of allowed operations, they are far more numerous. In addition, any attempt to designate "rules" that must be followed at all scales of biological computation should heed that biological computation does not have to "follow" them, but repurpose existing variation to grow out of the problem.

In summary, evolution, development and neural plasticity are not natural kinds but are part of an expansive process of continual organic "unfolding" (Hiesinger, 2021), of *descent with modification*. Despite being separated by the two tiers of constraints (Figure 1), the characteristics of biological computation—that it is displaced, reflexive, and groundless—are scale-invariant. Therefore, even though the brain is not as computationally powerful as evolution or development, which involve fuller senses of descent with modification, the holy trinity of symbol—memory—processing-structure is collapsed in biological computation. Nonetheless, even given the context-dependence of living systems, I take skepticism at the project of defining biological computation to be premature. Before spelling out the implications of this view for attempts to ground representational vehicles in neural activity, I critique this focus on activity first through presenting a recent case study from experimental systems neuroscience.

Section 3 — Implications of optogenetics and biological computation for representation and memory trace

In Section 3.1, I use a case study—Ortega-de San Luis *et al*. (2023)—to demonstrate that relevant causal variables that are not neural activity—and which are relevant to issues of grounding representations—can be identified through a combination of microscopy and optogenetic causal interventionist methods (à la Woodward, 2003). This argument from empirical research on its own challenges the status quo assumption that the neural vehicles of representation are exclusively to be grounded in neural activity, however interpreting these experimental results through biological computation will helps us contextualise them for how radical they are for changing our focus on enrol activity. Although this will damage assumptions about representation in neural systems, the concept of memory trace that informs the optogenetics studies *will not be saved either,* as the theory of biological computation will force us to reinterpret this literature as well. Ultimately I will conclude that these optogenetic technologies are key to broadening the discussion on neural representation, but that the concept of the engram or memory trace when treated ontologically is not. This frees up room for an excitingly fresh landscape for new theorising in neuroscience and philosophy of neuroscience, one that tacks closer to biology as opposed to computer science, logic, and mathematics.

3.1 — Reinterpreting attempts to ground representational vehicles in neural activity through the lens of optogenetics

Ortega-de San Luis *et al*. (2023) "*Engram cell connectivity as a mechanism for information encoding and memory function*" corroborates the hypothesis that alterations to connectivity patterns are one basis of learning (Chklovskii, Mel & Svoboda, 2004; Ryan *et al*., 2021). In other words, differences in connectivity patterns as a result of learning can explain differences in memory behaviour. Using a classic mouse-optogenetics paradigm, Ortega-de San Luis *et al*. (2023)

looked at changes in wiring in response to a specific experience and tested the functionality of these connectivity changes. First, mice formed two different associative memories (neutral-context memory A, and fear-memory B), and were then introduced to a shock while in context A, updating this association of A to A'. Looking at monosynaptic connections from hippocampal ventral CA1 (vCA1) to the basolateral amygdala (BLA), this updating process recruited some of the cells for fear memory B, while in a control condition—not introducing a shock during re-exposure to context A—did not result in overlap. The functionality of these connectivity changes was tested using optogenetic activation and inhibition protocols. First, vCA1 cells that were active while the animals were in in context A were tagged using a retrovirus, while the BLA neurons active in context B were tagged using a genetic switch. Later, reactivating vCA1 neurons tagged in context A, but while the animal was not in context A, was sufficient to increase freezing behaviour and involved downstream reactivation of the postsynaptic BLA neurons tagged during context B. This was not observed in control animals that had vCA1 neurons tagged in context A but which did not undergo the fear updating of context A to A'. Then, to test the necessity of the postsynaptic context B neurons in the BLA in those mice that had undergone updating, these neurons were inhibited while the mice were placed back in context A. This resulted in a drop in freezing behaviour, not observed in a control condition where by random BLA neurons were silenced. Further analyses of the membrane proteins involved in making these connections was performed, but for our purposes here we need only note the following: these changes in synaptic wiring were needed to alter the "content" of the associative memory.

In other words, what explains the "content" of the memory here is a non-activity neural variable or parameter, specifically a pattern of connectivity rather than of activity. Data collected from electrophysiological methods (e.g. analysis of spike trains) do not feature the explanans. In this context, activity may be compared to mere water, the shape through which it is guided determining what it can do, but which cannot "do" on its own, just as the connectivity pattern cannot migrate to the motor cortex or muscles to exert an effect. This study is also a classic example of how brain structure changes with learning and how these changes make

meaningful differences for the organism. What are the consequences of this for issues about neural representation? This study highlights how a completely different style of systems and molecular neuroscience can be used to explain behaviours that may involve internal representation. However, I envisage two conservative reactions:

Option 1 is to object to the prospect of expanding the causal variables functioning as representational vehicles beyond electrophysiological patterns to other features of the nervous system on the grounds that this opens up a potentially exponential and endless antecedent causal chain (e.g. why stop at using connectomic changes to explain, why not go back all the way to the past event or evolutionary history of the species? The argument may go that bounding criteria on our levels of explanation are needed, and since the final causal link or "driver" of behaviour is populations of neurons firing,[63] it is best not to involve anything else in our explanations of representations for the sake of explanatory parsimony. It could claim that all non-activity variables are at best "2nd-level explainers" (Barak and Krakauer, 2021). However, there is a second, weaker and seemingly more productive version of option 1 that *does* allow non-activity variables into attempts to explain representations.

Option 2 is to still hold that representations are to be grounded in neural activity, but non-activity parameters could *explain* in cases where past events exert a determining force on representational content (through, for example, identifiable intermediate changes in the connectome). Defining this view more clearly, grounding representations in neural activity would constitute the domain of the *representational status question*, while non-activity neural variables could be trafficked into explanations for the *content determination question* in cases where neural activity alone is explanatorily insufficient (which may be all cases).[64] This could prompt new questions about how the neuroscience drawn from should be

[63] However, to say that "firing neurons drive behaviour" is not necessarily to offer and adequate causal explanation of behaviour (Potter and Mitchell, 2024).

[64] The most liberal version of Option 2 would allow non-activity variables into explanations of both the representational status question and the content determination question. However, it would never attempt to explain the representational status question without neural activity featuring in the explanation.

apportioned. Could there be explanations of the content determination question that do not involve neural activity at all? Is neural activity alone sufficient to explain the representational status question, or are non-neural variables necessary?

I do not think that these are the most productive questions in the final analysis, and offer several objections to the strategy of Option 2, both immediate and from the view biological computation articulated above.

The first objection is that determining what does and does not count as cases where there is influence from the past is fraught. If perception is regarded as an inferential computation, and all inference involves the past ("priors"), then even the most stripped back tasks that could be considered in thought experiments must include past influences. This strategy could be saved by admitting that all cases of attempts to ground the content determination question will have to go beyond neural activity in their explanations. This is a huge and uncharted project.[65] Ultimately, both option 1 and option 2 could be construed as an attempt to preserve a classical view of computation. They ignore the basic causal topology of biological computation advanced in Section 2 and miss the key insight of biological computation, that neural activity is not this sufficiently independent substance that can be used in explanations because this is not how brains compute.

3.2 — Implications of biological computation for representation

The view of biological computation articulated in Section 2 would disagree with limiting the representational status question to neural activity at the very least (whatever about the content determination question), because from this perspective the biological mechanisms cited cannot be split into "patterns in

[65] Engaging with the emerging literature on how we understand causality in biology (Ross, 2023) may be of benefit. For example, classifying circuit wiring motifs as causal constraints would not prevent us using them to explain how variation in putative representational vehicles influences the determination of representational content. This could borrow loosely from Dretske's (1988) differentiation between "structuring causes" and "triggering causes" of behaviour, mapping them to connectome and activity, respectively. As this is not a philosophy of causation paper, this connection only suggested for those interested.

spikes" and "everything else," at least not if we want our explanations to work. For anyone who finds the view of biological computation disagreeable, Option 2 may be a fruitful area for further development of the existing philosophical literature on representation that does not challenge the core assumptions that make it possible.

To claim that biological systems operate on themselves is not necessarily to claim that these operations are classically "self-referential." Issues of "reference" pertain typically to the realm of human linguistic and artistic reference (words, symbols and gestures) and to systems designed by humans who communally agree on reference systems to use. I have argued that biological systems do not employ a fundamental abstraction between processes and symbols. To draw out the implications of this for issues about representation, consider Goodman's differentiation between representation and resemblance: "A Constable painting of Marlborough Castle is more like any other picture than it is like the Castle, yet it represents the Castle and not another picture" (Goodman, 1968, p.5). Note that in this instance we are dealing crucially with two separate objects—painting and castle. The brain is not two separate objects, and *this is what was meant by "the brain as twofold object*." Brain structure and neural activity patterns are not two different objects, neither has to represent the other. They are of one and the same computational process. Neither do they together form a hybrid computational system that is a canvas in search of something called "the world" that must be "represented." This is contra to any positing of representations as "internal states," as intermediaries between inputs and outputs that internalise a copy of external statistical regularities. Self-computation means we are not dealing with relations between two objects, and that to frame biological computation in terms of externality-internality is to contort it to suit a paradigm designed for manipulating symbols on paper and running syllogisms on semiconductor chips.

Biological computation also clearly complicates attempts to ground vehicles of representation in "neural activity," which can now be understood as an intermediary in a process of self-operation. Grounding content, to the extent that this is still deemed a sensible philosophical project to pursue, may require more

than just considering “both activity and structure” but an understanding of how computational processes traverse structure and activity and back. This is a good moment to link this discussion of memory traces.

3.3— Implications of biological computation for memory traces a la “stored information”

Naturalising the concept of the memory trace is another project in neuroscience and in the emerging field of philosophy of memory, paralleling attempts to ground representations. Like the latter, memory traces are explanatory posits—they are needed to be able to explain behaviours in which an organism responds in a manner that could only have been possible if “learning” from past events had persisted. There are a variety of theories in the philosophy of memory about how to understand whether traces are needed and what they should be understood as (De Brigard, 2014; Robins, 2020). I wish to focus here instead on distinguishing between two different interpretations of the memory traces or “engram" concept—one is a pragmatic reference to cells involved in learning, the other is an ontological posit—that there exists “stored information.” The meaning of the term engram moonlights interchangeably in the same papers, (Josselyn, Köhler & Frankland, 2015; Josselyn & Tonegawa, 2020; Ryan *et al*., 2021). I argue against the latter because of the naive realism about information that it entails is incompatible with biological computation.

The first meaning of engram is practical: it now refers to a collection of cells active during learning and are necessary for recall behaviour. This is a pragmatic operationalisation of the old idea of a memory trace: that there is something left over by learning which is needed for remembering. The ontological interpretation takes “the engram” concept to imply the existence of “information” out there in the brain that is stored. This ontological interpretation of the engram is a form of naive realism: there really are encodings of memories as information etched into the brain somehow.

Mace (2021) highlights that there are other operationalisations of the term “engram” that do not obviously imply a conception of information storage, such as Mayford (2014)—“the cellular and molecular change in the brain that occurs with learning and alters the subsequent processing of specific external cues to produce memory recall.” This “alteration of subsequent processing” fits with the intuition of memory as descent with modification. Recall from Section 2.3.i that “activity patterns and sequences can induce changes in the circuits themselves. This changes the very structure that previously “processed” them, leading to different results in future." This would also explain why there has been no success regarding memory trace decoding. Because the organism *is its past*, and does not “store” it, one needs a way not to “decode” but to “unravel” the network, and this is at the very least practically impossible.

From the perspective on biological computation, we can appreciate how naive realism about memory traces as entities in the brain to be decoded is flawed by taking a comparative approach to the different systems in which hysteresis occurs. Consider the following: A trail left by a snail is not a snail, or even an encoding of a snail’s velocity. Yet, when downstream products of present neural processes are retained in the nervous system, these products have far more potential on account of the kind of system in which they are embedded (a trace in the brain is not a trace on the ground, at least not all of them). So what is the crucial difference and where does it come from? Although a snail leaves a record of location left over by movement, and could be called a trace, the potential record has been externalised as a separate object. To use it as a trace would require a completely different process, that of an organism evolving the sensory competencies for the excreted products of past actions to feature semiotically in its umwelt as potentially useful affordances. This applies across the boundary of individuality within organisms too, for example in neuronal development neuronal filopodia follow the chemical trails left behind by neurons that have migrated before them to a given area (Wit & Hiesinger, 2023). Here too past action has been externalised. In contrast, the principle of biological computation I have considered above involves displaced operation on self. This reflexive usability separates the products of neural

processes from scars and sunburns, which cannot feed back to influence neural processes.

This can explain why there is no symbolic “information” or “data” to be found in the nervous system: it never needed to invent such a thing to begin with. In one way this is foreign to what we are able to conceive because of the standards set by the richly constructed symbolism of our daily lives (Baudrillard, 1981/1994), yet on a phenomenological level we do not have to “search” for in our minds eye to recall something (even if this aids non-aphantasics). Thus, if traces are to be “saved” as a concept useful to neuroscience, all usage of “encoding” or “encoding information” must be dropped. In addition, as we are abandoning a naive realist interpretation of memory traces, then any idea that traces are things that remain in connectivity structures must also go, as should should the idea that traces are universals that transmigrate through a variety of biological substrates (e.g. from activity to connectome to activity and back again, e.g. through the cycles typically denoted learning, consolidation, recall and reconsolidation). The only way to satisfy this is to dispense with “trace” as a noun that implies things to look for, for “trace” as a verb—to trace though. Even still, “trace” can imply resemblance or representation to something else, i.e. thereby framing a problem in terms of two objects. The same critique about the brain not being the kind of system that “refers” to other objects applies to memory traces. Ultimately, from the perspective of biological computation, “information” in the neuroscientific literature should be treated as an empty signifier that functions as part of a pragmatist epistemology unlikely to change any time soon, but should not interpreted as an ontological category.

Section 4 — Concluding remarks

Attempting to naturalise the concept of mental representations in terms of neural representations requires the broader demands of a biologically respectable view of neural computation. I have shown that existing theories of computation are not applicable to the brain or any biological system, and offered a best shot at

explaining computation in living systems—biological computation. The defining features of biological computation are reflexivity, displacement, and groundlessness. Despite this new framework, we still fall short of naturalising mental representations. I have located this shortcoming in the near-exclusive focus on neural activity at the expense of other areas of neuroscience, and used my account of biological computation to argue that neural activity should not be singled out as a level of organisation, and one that possesses primary explanatory and causal privilege at that. Although my argument is agnostic about representations, it is not incompatible with vehicle realism. For anyone committed to vehicle realism, however, there can be no talk of computation over representation abstract from the materiality of biology, and given this inherent medium dependence, it is not enough to be looking at only one part of the medium and tell a story of computation that excludes the rest when there is evidence it matters. For all the criticism of classical computation and the focus on neural activity, modern computational neuroscience does place an important emphasis on processes and dynamics. Biological computation does not seek to disregard this work, but show that the remainder of the biology *is the dynamics too*, rather than fixed structures that "implement" dynamics and can therefore be relegated to the background. Defining biological computation formally as was done by Turing for digital computation will have to wait for another time, the very act of representing it on paper may prove problematic. Getting to human abstraction will be even more challenging, since we cannot look to neural activity alone for simple symbolic primitives. Looking for alternatives to operator-operand constructions to structure our thinking may require nothing less than rewriting the rules set theory for a mathematics suitable to describe biological computation. Unpacking this may require a philosophical revolution in neuroscience on par with Peter Mitchell's chemiosmosis or Lynn Margulis' endosymbiosis in general biology—not just drastic shifts in worldviews, but drastic shifts in how we understand life and living.

Thesis Conclusion

The deluge of papers and perspectives published weekly in experimental and computational neuroscience is daunting. The explosion of methods in the preceding decades has been and will continue to drive groundbreaking research. The ideas at the heart of neuroscience in contrast change far more slowly. It can take years to intuit what they even are—the categories of our cognitive ontology and of information, computation, and representation—let alone begin to tease out how they structure scientific theorising and to challenge them. Neuroscientific theory remains in its infancy. The ultimate motivation of this thesis was to make small increments in the realm of how we understand memory and information storage. In employing a methodology of taking an under-theorised area of neuroscience as my focus point—that of the emerging field of engram neuroscience—and bringing it into dialogue with various different disciplines, I have established new bridges between memory neuroscience and philosophy, memory neuroscience and machine-learning, and theories of computation in biology. This thesis sets the stage for more rigorously formal work by philosophers, computer scientists, AI researchers, and mathematicians, such new computational models of memory and attempts to naturalise mental representations and memory traces. Undertaking this project will lead to direct confrontation with the core assumptions of our time. Far from relegation to a secondary concern, critical reflection and interrogation of memory is a viable and unfinished track into the core foundational issues in the cognitive sciences.

Bibliography

Abdou, K., Shehata, M., Choko, K., Nishizono, H., Matsuo, M., Muramatsu, S. I., & Inokuchi, K. (2018). Synapse-specific representation of the identity of overlapping memory engrams. *Science (New York, N.Y.)*, *360*(6394), 1227–1231. https://doi.org/10.1126/science.aat3810

Anderson D. J. (2016). Circuit modules linking internal states and social behaviour in flies and mice. *Nature reviews. Neuroscience*, *17*(11), 692–704. https://doi.org/10.1038/nrn.2016.125

Autore, L., O'Leary, J. D., Ortega-de San Luis, C., & Ryan, T. J. (2023). Adaptive expression of engrams by retroactive interference. *Cell reports*, *42*(8), 112999. https://doi.org/10.1016/j.celrep.2023.112999

Attardo, A., Fitzgerald, J. E., & Schnitzer, M. J. (2015). Impermanence of dendritic spines in live adult CA1 hippocampus. Nature, 523(7562), 592–596. https://doi.org/10.1038/nature14467

Avery, O.T., MacLeod, C.M., McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *Journal of Experimental Medicine*. **79** (2): 137–158. doi:10.1084/jem.79.2.137.

Barack, D.L., Krakauer, J.W. (2021). Two views on the cognitive brain. *Nat Rev Neurosci* **22**, 359–371. https://doi.org/10.1038/s41583-021-00448-6

Barron, H. C., Vogels, T. P., Behrens, T. E., & Ramaswami, M. (2017). Inhibitory engrams in perception and memory. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(26), 6666–6674. https://doi.org/10.1073/pnas.1701812114

Baudrillard, J. (1994). Simulacra and simulation (S. F. Glaser, Trans.). University of Michigan. (Original work published 1981)

Bédécarrats, A., Chen, S., Pearce, K., Cai, D., & Glanzman, D. L. (2018). RNA from Trained *Aplysia* Can Induce an Epigenetic Engram for Long-Term Sensitization in Untrained *Aplysia*. *eNeuro*, *5*(3), ENEURO.0038-18.2018. https://doi.org/10.1523/ENEURO.0038-18.2018

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954–967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Bliss, T. V., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of physiology*, *232*(2), 331–356. https://doi.org/10.1113/jphysiol.1973.sp010273

Bolsius, Y. G., Heckman, P. R. A., Paraciani, C., Wilhelm, S., Raven, F., Meijer, E. L., Kas, M. J. H., Ramirez, S., Meerlo, P., & Havekes, R. (2023). Recovering object-location memories after sleep deprivation-induced amnesia. *Current biology : CB*, *33*(2), 298–308.e5. https://doi.org/10.1016/j.cub.2022.12.006

Brette, R. (2019) 'Is coding a relevant metaphor for the brain?', *Behavioral and Brain Sciences*. 2018/07/16. Cambridge University Press, 42, p. e215. DOI: 10.1017/ S0140525X19000049.

Brosens, N., Lesuis, S. L., Bassie, I., Reyes, L., Gajadien, P., Lucassen, P. J., & Krugers, H. J. (2023). Elevated corticosterone after fear learning impairs remote

auditory memory retrieval and alters brain network connectivity. *Learning & memory (Cold Spring Harbor, N.Y.)*, *30*(7), 125–132. https://doi.org/10.1101/lm.053836.123

Burnston, D.C. (2021). Contents, vehicles, and complex data analysis in neuroscience. *Synthese* 199 (1-2):1617-1639.

Buzsáski G. (2006). *Rhythms of the Brain*. Oxford University Press.

Buzsáki G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, *68*(3), 362–385. https://doi.org/10.1016/j.neuron.2010.09.023

Callard, F., & Fitzgerald, D. (2015). *Rethinking Interdisciplinarity across the Social Sciences and Neurosciences*. Palgrave Macmillan.

Campbell, R. R., & Wood, M. A. (2019). How the epigenome integrates information and reshapes the synapse. *Nature reviews. Neuroscience*, *20*(3), 133–147. https://doi.org/10.1038/s41583-019-0121-9

Cao, R. (2012). A Teleosemantic approach to information in the brain. *Biology and Philosophy*.

Cao, R. (2022). Putting representations to use. *Synthese* 200 (2).

Cao, R., & Rathkopf, C. (2019). Modest and immodest neural codes: Can there be modest codes?. *The Behavioral and brain sciences*, *42*, e221. https://doi.org/10.1017/S0140525X19001420

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature reviews. Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Cisek, P. (2019) ‘Resynthesizing behavior through phylogenetic refinement’, *Attention, Perception, & Psychophysics*, 81(7), pp. 2265–2287. doi: 10.3758/s13414-019-01760-1.

Chandra, S., Sharma, S., Chaudhuri, R., & Fiete, I.R. (2024). High-capacity flexible hippocampal associative and episodic memory enabled by prestructured “spatial” representations. bioRxiv.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress.* New York: OUP USA.

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013–1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Chapman, D. P., Power, S. D., Vicini, S., Ryan, T. J., & Burns, M. P. (2024). Amnesia after Repeated Head Impact Is Caused by Impaired Synaptic Plasticity in the Memory Engram. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *44*(8), e1560232024. https://doi.org/10.1523/JNEUROSCI.1560-23.2024

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., & Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature neuroscience*, *22*(9), 1512–1520. https://doi.org/10.1038/s41593-019-0460-x

Chen, D., Lou, Q., Song, X. J., Kang, F., Liu, A., Zheng, C., Li, Y., Wang, D., Qun, S., Zhang, Z., Cao, P., & Jin, Y. (2024). Microglia govern the extinction of acute stress-induced anxiety-like behaviors in male mice. *Nature communications*, *15*(1), 449. https://doi.org/10.1038/s41467-024-44704-6

Chen, S., Cai, D., Pearce, K., Sun, P. Y.-W., Roberts, A. C., and Glanzman, D. L. (2014). Reinstatement of long-term memory following erasure of its behavioral and synaptic expression in Aplysia. eLife 3:e03896. doi: 10.7554/eLife.03896

Chirimutra, M. (2024). *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience.* MIT Press.

Chklovskii, D. B., Mel, B. W., & Svoboda, K. (2004). Cortical rewiring and information storage. *Nature*, *431*(7010), 782–788. https://doi.org/10.1038/nature03012

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Clark A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cobb M. (2020). *The Idea of the Brain*. Profile Books

Colaço D, Najenson J. (2023). Where memory resides: Is there a rivalry between molecular and synaptic models of memory? *Philosophy of Science*.1-11. doi:10.1017/psa.2023.126

Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., Salzman, C. D., Mamelak, A. N., Fusi, S., & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, *632*(8026), 841–849. https://doi.org/10.1038/s41586-024-07799-x

Craver, C. (2007). *Explaining the Brain*. Oxford University Press.

Czégel, D., Giaffar, H., Tenenbaum, J. B., & Szathmáry, E. (2022). Bayes and Darwin: How replicator populations implement Bayesian computations.

BioEssays : news and reviews in molecular, cellular and developmental biology, *44*(4), e2100255. https://doi.org/10.1002/bies.202100255

Dasgupta, I., & Gershman, S. J. (2021). Memory as a Computational Resource. *Trends in cognitive sciences*, *25*(3), 240–251. https://doi.org/10.1016/j.tics.2020.12.008

Deacon, T. W. (2021). How molecules became signs. *Biosemiotics*, **14**, 537–559. https://doi.org/10.1007/s12304-021-09453-9

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, 191, 155-185.

D'Elia, T., Veerapaneni, R., Theraisnathan, V., & Rogers, S. O. (2009). Isolation of fungi from Lake Vostok accretion ice. *Mycologia*, *101*(6), 751–763. https://doi.org/10.3852/08-184

de Kloet, E. R., & Joëls, M. (2024). The cortisol switch between vulnerability and resilience. *Molecular psychiatry*, *29*(1), 20–34. https://doi.org/10.1038/s41380-022-01934-8

de Quervain, D., Schwabe, L., & Roozendaal, B. (2017). Stress, glucocorticoids and memory: implications for treating fear-related disorders. *Nature reviews. Neuroscience*, *18*(1), 7–19. https://doi.org/10.1038/nrn.2016.155

Dennet D. C. (1978). Toward a Cognitive Theory of Consciousness, in C.W. Savage, ed., Minnesota Studies in the Philosophy of Science, Volume 9: Perception and Cognitions, Issues in the Foundation of Psychology. Minneapolis: University of Minnesota Press, pp201-228.

Dijkstra, N., Kok, P., & Fleming, S. M. (2022). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience and*

biobehavioral reviews, *135*, 104557. https://doi.org/10.1016/j.neubiorev.2022.104557

Donald, M. (1991). Origins of the modern mind. Harvard University Press.

Donohoe, J. (2010). Man as Machine. Review of Memory and the Computational Brain, by C. R. Gallistel and A. P. King. *Behavior and Philosophy* 38:83–101.

Dorst, K., Senne, R.A., Diep, A.H., de Boer, A.R., Suthard, R.L., Leblanc, H., Ruesch, E.A. Skelton, S., McKissick, O.P., Bladon, J.H., Ramirez, S. (2023). *bioRxiv.* 2023.02.24.529744; https://doi.org/10.1101/2023.02.24.529744

Dragoi G. (2024). The generative grammar of the brain: a critique of internally generated representations. *Nature reviews. Neuroscience*, *25*(1), 60–75. https://doi.org/10.1038/s41583-023-00763-0

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dudai, Y. Memory: It's all about representations. In H.L. Roediger, Y. Dudai & S.M. Fitzpatrick (eds.) *Science of Memory: Concepts* (pp. 17-21). Oxford University Press.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011

Eichenbaum H. (2016). Still searching for the engram. *Learning & behavior*, *44*(3), 209–222. https://doi.org/10.3758/s13420-016-0218-1

Feyerabend, P. K. (1975). *Against Method*. New York: Verso Books

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon Press

Fodor J.A. (1981): 'The Mind-Body Problem', Scientific American 2 (January 1981). Reprinted in Heil, J. (ed.) (2004a). Philosophy of Mind: Guide and Anthology, Oxford: Oxford University Press

Friedrichs, J., & Kratochwil, F. (2009). On Acting and Knowing. How Pragmatism Can Advance International Relations Research and Methodology. International Organization, 63, 701-731. https://doi.org/10.1017/S0020818309990142

Friston, K. (2010) 'The free-energy principle: a unified brain theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138. doi: 10.1038/nrn2787.

Furber S. (2016). Large-scale neuromorphic computing systems. *Journal of neural engineering*, *13*(5), 051001. https://doi.org/10.1088/1741-2560/13/5/051001

Goode, T. D., Tanaka, K. Z., Sahay, A., & McHugh, T. J. (2020). An Integrated Index: Engrams, Place Cells, and Hippocampal Memory. *Neuron*, *107*(5), 805–820. https://doi.org/10.1016/j.neuron.2020.07.011

Gallistel C. R. (2017). The Coding Question. *Trends in cognitive sciences*, *21*(7), 498–508. https://doi.org/10.1016/j.tics.2017.04.012

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience.* Wiley-Blackwell.

Gazzaniga, M. (2018). *The Consciousness Instinct*. Farrar, Straus and Giroux.

Gershman, S.J., & Goodman, N.D. (2014). Amortized Inference in Probabilistic Reasoning. Cognitive Science, 36.

Gershman, S.J. (2023). The molecular memory code and synaptic plasticity: a synthesis. *BioSystems, 224*, 104825.

Godfrey-Smith, P. (2014). Sender-Receiver Systems within and between Organisms. *Philosophy of Science*, *81*(5), 866–878. https://doi.org/10.1086/677686

Gould, S., & Vrba, E. (1982). Exaptation—a Missing Term in the Science of Form. *Paleobiology, 8*(1), 4-15. doi:10.1017/S0094837300004310

Grothe B. (2003). New roles for synaptic inhibition in sound localization. *Nature reviews. Neuroscience*, *4*(7), 540–550. https://doi.org/10.1038/nrn1136

Guskjolen, A., & Cembrowski, M. S. (2023). Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular psychiatry*, 10.1038/s41380-023-02137-5. Advance online publication. https://doi.org/10.1038/s41380-023-02137-5

Han, W., Tellez, L. A., Rangel, M. J., Jr, Motta, S. C., Zhang, X., Perez, I. O., Canteras, N. S., Shammah-Lagnado, S. J., van den Pol, A. N., & de Araujo, I. E. (2017). Integrated Control of Predatory Hunting by the Central Nucleus of the Amygdala. *Cell*, *168*(1-2), 311–324.e18. https://doi.org/10.1016/j.cell.2016.12.027

Hardt, O., & Sossin, W. S. (2020). Terminological and Epistemological Issues in Current Memory Research. *Frontiers in molecular neuroscience*, *12*, 336. https://doi.org/10.3389/fnmol.2019.00336

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hebb, D.O. (1949). *The Organization of Behavior*. New York: Wiley

Hershey A, Chase M (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol*. 36 (1): 39–56. doi:10.1085/jgp.36.1.39.

Hiesinger, P.R. (2021). *The Self-Assembling Brain*. Princeton University Press.

Hoffmeyer, J. (2008). Semiotic Scaffolding Of Living Systems. In: Barbieri, M. (eds) Introduction to Biosemiotics. Springer, Dordrecht.

Hurley, S.L. (1998). *Consciousness in Action*. Harvard University Press.

Ishii, K. K., Osakada, T., Mori, H., Miyasaka, N., Yoshihara, Y., Miyamichi, K., & Touhara, K. (2017). A Labeled-Line Neural Circuit for Pheromone-Mediated Sexual Behaviors in Mice. *Neuron*, *95*(1), 123–137.e8. https://doi.org/10.1016/j.neuron.2017.05.038

Jaeger, H. (2021). Towards a generalized theory comprising digital, neuromorphic and unconventional computing. Neuromorphic Computing and Engineering, 1. https://doi.org/10.1088/2634-4386/abf151

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor?. *PLoS computational biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M., & Rettberg, P. (2008). Tardigrades survive exposure to space in low Earth orbit. *Current biology : CB*, *18*(17), R729–R731. https://doi.org/10.1016/j.cub.2008.06.048

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. *Nature reviews. Neuroscience*, *16*(9), 521–534. https://doi.org/10.1038/nrn4000

Josselyn, S. A., & Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. *Science (New York, N.Y.)*, *367*(6473), eaaw4325. https://doi.org/10.1126/science.aaw4325

Kandel E. R. (2001). The molecular biology of memory storage: a dialogue between genes and synapses. *Science (New York, N.Y.)*, *294*(5544), 1030–1038. https://doi.org/10.1126/science.1067020

Kappel, D., Legenstein, R., Habenschuss, S., Hsieh, M., & Maass, W. (2018). A Dynamic Connectome Supports the Emergence of Stable Computational Function of Neural Circuits through Reward-Based Learning. *eNeuro*, *5*(2), ENEURO.0301-17.2018. https://doi.org/10.1523/ENEURO.0301-17.2018

Kauffman, S. (2019). *A World Beyond Physics*. Oxford University Press.

Kay L. E. (2001). From logical neurons to poetic embodiments of mind: Warren S. McCulloch's project in neuroscience. *Science in context*, *14*(4), 591–614. https://doi.org/10.1017/s0269889701000266

Kazanina, N., & Poeppel, D. (2023). The neural ingredients for a language of thought are available. *Trends in cognitive sciences*, *27*(11), 996–1007. https://doi.org/10.1016/j.tics.2023.07.012

Kaznatcheev, A., & Kording, K.P. (2022). Nothing makes sense in deep learning, except in the light of evolution. ArXiv, abs/2205.10320.

Kelty-Stephen, D.G., Cisek, P.E., Bari, B.D., Dixon, J., Favela, L.H., Hasselman, F., Keijzer, F.A., Raja, V., Wagman, J.B., Thomas, B.J., & Mangalam, M. (2022).

In search for an alternative to the computer metaphor of the mind and brain. https://doi.org/10.48550/arXiv.2206.04603

Koch, C. (1999) Biophysics of Computation: Information Processing in Single Neurons. New York: Oxford University Press

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kirschner, M., & Gerhart, J. (1998). Evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8420–8427. https://doi.org/10.1073/pnas.95.15.8420

Kitamura, T., Ogawa, S. K., Roy, D. S., Okuyama, T., Morrissey, M. D., Smith, L. M., Redondo, R. L., & Tonegawa, S. (2017). Engrams and circuits crucial for systems consolidation of a memory. *Science (New York, N.Y.)*, *356*(6333), 73–78. https://doi.org/10.1126/science.aam6808

Kol, A., & Goshen, I. (2021). The memory orchestra: the role of astrocytes and oligodendrocytes in parallel to neurons. *Current opinion in neurobiology*, *67*, 131–137. https://doi.org/10.1016/j.conb.2020.10.022

Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface focus*, *8*(6), 20180041. https://doi.org/10.1098/rsfs.2018.0041

Koolschijn, R. S., Emir, U. E., Pantelides, A. C., Nili, H., Behrens, T. E. J., & Barron, H. C. (2019). The Hippocampus and Neocortical Inhibitory Engrams Protect against Memory Interference. *Neuron*, *101*(3), 528–541.e6. https://doi.org/10.1016/j.neuron.2018.11.042

Koonin E. V. (2016). Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC biology, 14(1), 114. https://doi.org/10.1186/s12915-016-0338-2

Koren, T., Yifa, R., Amer, M., Krot, M., Boshnak, N., Ben-Shaanan, T. L., Azulay-Debby, H., Zalayat, I., Avishai, E., Hajjo, H., Schiller, M., Haykin, H., Korin, B., Farfara, D., Hakim, F., Kobiler, O., Rosenblum, K., & Rolls, A. (2021). Insular cortex neurons encode and retrieve specific immune responses. *Cell*, *184*(24), 5902–5915.e17.

Koren, T., & Rolls, A. (2022). Immunoception: Defining brain-regulated immunity. *Neuron*, *110*(21), 3425–3428. https://doi.org/10.1016/j.neuron.2022.10.016

Krakauer, D.C., Bertschinger, N., Olbrich, E., Flack, J. C., & Ay, N. (2020). The information theory of individuality. *Theory in biosciences*, *139*(2), 209–223. https://doi.org/10.1007/s12064-020-00313-7

Krakauer, D.C. & Kempes, C. (2024, September 17) Problem-solving matter. *Aeon*. https://aeon.co/essays/is-life-a-complex-computational-process

Krakauer, J.W. (2022). Representation in Cognitive Science by Nicholas Shea: But Is It Thinking? The Philosophy of Representation Meets Systems Neuroscience. *Studies in History and Philosophy of Science*, *92*, 267-269. https://doi.org/10.1016/j.shpsa.2021.05.014.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Kumaran, D., Hassabis, D., & McClelland, J.L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends in Cognitive Sciences, 20, 512-534.

Langdon, C., Genkin, M., & Engel, T. A. (2023). A unifying perspective on neural manifolds and circuits for cognition. *Nature reviews. Neuroscience*, *24*(6), 363–377. https://doi.org/10.1038/s41583-023-00693-x

Laplane, L. (2016). *Cancer Stem Cells: Philosophy and Therapies*. Harvard University Press.

Latour, B. & Woolgar, S. (1987). Laboratory Life: The Construction of Scientific Facts. Princeton University Press.

Lee, J. (2019). Structural representation and the two problems of content. *Mind and Language,* 34 (5):606-626. https://doi.org/10.1111/mila.12224

Lehtinen S. O. (2021). Ecological and evolutionary consequences of predator-prey role reversal: Allee effect and catastrophic predator extinction. *Journal of theoretical biology*, *510*, 110542. https://doi.org/10.1016/j.jtbi.2020.110542

Lei, B., Lv, L., Hu, S., Tang, Y., & Zhong, Y. (2022). Social experiences switch states of memory engrams through regulating hippocampal Rac1 activity. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(15), e2116844119. https://doi.org/10.1073/pnas.2116844119

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv*, abs/2005.11401

Lewontin, R. C. (1983). The Organism as the Subject and Object of Evolution. In R. Levins and R.C. Lewontin (eds.), *The Dialectical Biologist*. Harvard University Press.

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., & Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, *484*(7394), 381–385. https://doi.org/10.1038/nature11028

Luo, L. (2020). *Principles of Neurobiology*. (2nd ed). Taylor & Francis.

Lynch G. (1998). Memory and the brain: unexpected chemistries and a new pharmacology. *Neurobiology of learning and memory*, *70*(1-2), 82–100. https://doi.org/10.1006/nlme.1998.3840

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737–749. https://doi.org/10.1016/j.neuron.2011.07.012

Mace, C. B. (2021). *Casting light on the search for engrams: On the reductionism-mechanism debate* (Order No. 28644486). Available from ProQuest Dissertations & Theses A&I; Publicly Available Content Database. (2597766427).

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *155*(1), 117–131. http://www.jstor.org/stable/41487690

Maley, C. J. (2021). The physicality of representation. *Synthese* 199 (5-6):14725-14750 https://doi.org/10.1007/s11229-021-03441-9#Fn1

Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., & Wu, J. (2019). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. ArXiv, abs/1904.12584.

Marr, D. (1982) *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, Massachusetts: MIT Press.

Martin, S. J., Grimwood, P. D., & Morris, R. G. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual review of neuroscience*, *23*, 649–711. https://doi.org/10.1146/annurev.neuro.23.1.649

Maturana, H. R. & Varela F.J. (1991). *Autopoiesis and Cognition: The Realization of the Living* (3rd ed.). Springer Science & Business Media (Original work published 1972 as *De Maquinas y Seres Vivos* by Editorial Universitaria S.A)

Mayford, M. (2014). The Search for a Hippocampal Engram. *Philosophical Transactions of the Royal Society B* 369 (1): 1–10.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological review, 102 3, 419-457 .

McGaugh J. L. (2000). Memory--a century of consolidation. *Science (New York, N.Y.)*, *287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

McNaughton, B.L. (2010). Cortical hierarchies, sleep, and the extraction of knowledge from memory. Artif. Intell., 174, 205-214.

Mei, J., Muller, E., & Ramaswamy, S. (2022). Informing deep neural networks by multiscale principles of neuromodulatory systems. *Trends in neurosciences*, *45*(3), 237–250. https://doi.org/10.1016/j.tins.2021.12.008

Mermillod, M., Bugaiska, A., & Bonin, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. Frontiers in psychology, 4, 504. https://doi.org/10.3389/fpsyg.2013.00504

Miller K. D. (2016). Canonical computations of cerebral cortex. *Current opinion in neurobiology*, *37*, 75–84. https://doi.org/10.1016/j.conb.2016.01.008

Mitchell, K. (2023). The origins of meaning – from pragmatic control signals to semantic representations. *PsyArXiv*. Doi: 10.31234/osf.io/dfkrv

Mitchell, K. J., & Potter, H. (2024, June 26). Beyond mechanism – extending our concepts of causation in neuroscience. https://doi.org/10.31234/osf.io/63qvy

Moreno, A. & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer

Moskovitch, M. (2007). Memory: Why the engram is elusive. In H.L. Roediger, Y. Dudai & S.M. Fitzpatrick (eds.) *Science of Memory: Concepts* (pp. 17-21). Oxford University Press.

Najenson, J. (2021). What have we learned about the engram? *Synthese* **199**, 9581–9601. https://doi.org/10.1007/s11229-021-03216-2

Nambu, M. F., Lin, Y. J., Reuschenbach, J., & Tanaka, K. Z. (2022). What does engram encode?: Heterogeneous memory engrams for different aspects of experience. *Current opinion in neurobiology*, *75*, 102568. https://doi.org/10.1016/j.conb.2022.102568

Neander, K. (2017). *A Mark of the Mental: A Defence of Informational Teleosemantics*. Cambridge: MIT Press

Ohta T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, *246*(5428), 96–98. https://doi.org/10.1038/246096a0

O'Leary J., Bruckner, R., Autore, L., Ryan, T.J. (2023). Natural forgetting reversibly modulates engram expression. *eLife,* **12**:RP92860. https://doi.org/10.7554/eLife.92860.1

Orlandi, N. (2020). Representing as Coordinating with Absence. In J. Smortchkova, K. Dołęga, & T Schlicht (Eds.), *What are Mental Representations?* New York: OUP

Ortega-de San Luis, C., & Ryan, T. J. (2022). Understanding the physical basis of memory: Molecular mechanisms of the engram. *The Journal of biological chemistry*, *298*(5), 101866. https://doi.org/10.1016/j.jbc.2022.101866

Ortega-de San Luis, C., Pezzoli, M., Urrieta, E., Ryan, T.J. (2023). Engram cell connectivity as a mechanism for information encoding and memory function. *Current Biology,* DOI: https://doi.org/10.1016/j.cub.2023.10.074

Pastuzyn, E. D., Day, C. E., Kearns, R. B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., Yoder, N., Belnap, D. M., Erlendsson, S., Morado, D. R., Briggs, J. A. G., Feschotte, C., & Shepherd, J. D. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*, *172*(1-2), 275–288.e18. https://doi.org/10.1016/j.cell.2017.12.024

Pattee, H.H. (2012). Cell Psychology: an evolutionary approach to the symbol-matter problem. In H.H. Pattee & Rączaszek-Leonardi, J. (Ed.), *LAWS, LANGUAGE and LIFE: Howard Pattee's classic papers on the physics of symbols with contemporary commentary*. (Original work published 1987)

Perusini, J. N., Cajigas, S. A., Cohensedgh, O., Lim, S. C., Pavlova, I. P., Donaldson, Z. R., & Denny, C. A. (2017). Optogenetic stimulation of dentate gyrus engrams restores memory in Alzheimer's disease mice. *Hippocampus*, *27*(10), 1110–1122. https://doi.org/10.1002/hipo.22756

Piccinini, G. (2008). Computation without Representation. Philosophical Studies, 137, 205-241.

Piccinini, G. (2007). Computational modeling vs. computational explanation: Is everything a Turing machine, and does it matter to the philosophy of mind? Australasian Journal of Philosophy 85 (1):93 – 115.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive science*, *37*(3), 453–488. https://doi.org/10.1111/cogs.12012

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford: OUP

Poeppel, D., & Idsardi, W. (2022). We don't know how the brain stores anything, let alone words. *Trends in cognitive sciences*, *26*(12), 1054–1055. https://doi.org/10.1016/j.tics.2022.08.010

Poll, S., Mittag, M., Musacchio, F., Justus, L. C., Giovannetti, E. A., Steffen, J., Wagner, J., Zohren, L., Schoch, S., Schmidt, B., Jackson, W. S., Ehninger, D., & Fuhrmann, M. (2020). Memory trace interference impairs recall in a mouse model of Alzheimer's disease. *Nature neuroscience*, *23*(8), 952–958

Power, S., Stewart, E., Zielke, L., Byrne, E., Ortega-de San Luis, L., Lynch, L., Ryan, T. (2023). Immune activation state modulates infant engram expression across development. *Science Advances,* **9**,eadg9921 .DOI:10.1126/sciadv.adg9921

Queenan, B. N., Ryan, T. J., Gazzaniga, M. S., & Gallistel, C. R. (2017). On the research of time past: the hunt for the substrate of memory. *Annals of the New York Academy of Sciences*, *1396*(1), 108–125. https://doi.org/10.1111/nyas.13348

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *The Behavioral and brain sciences*, *46*, e261. https://doi.org/10.1017/S0140525X22002849

Ramsey, W.M. (2007). *Representation Reconsidered*. Cambridge University Press.

Ramirez, S., Liu, X., MacDonald, C. J., Moffa, A., Zhou, J., Redondo, R. L., & Tonegawa, S. (2015). Activating positive memory engrams suppresses depression-like behaviour. *Nature*, *522*(7556), 335–339. https://doi.org/10.1038/nature14514

Ratzon, A., Derdikman, D., & Barak, O. (2024). Representational drift as a result of implicit regularization. *eLife*, *12*, RP90069. https://doi.org/10.7554/eLife.90069

Rehling, J., & Hofstadter, D.R. (1997). The parallel terraced scan: an optimization for an agent-oriented architecture. 1997 IEEE International Conference on Intelligent Processing Systems (Cat. No.97TH8335), 1, 900-904 vol.1.

Reijmers, L. G., Perkins, B. L., Matsuo, N., & Mayford, M. (2007). Localization of a stable neural correlate of associative memory. *Science (New York, N.Y.)*, *317*(5842), 1230–1233. https://doi.org/10.1126/science.1143839

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Richards, B. A., & Kording, K. P. (2023). The study of plasticity has always been about gradients. *The Journal of physiology*, *601*(15), 3141–3149. https://doi.org/10.1113/JP282747

Robins, S. K. (2018). Memory and Optogenetic Intervention: Separating the Engram from the Ecphory. *Philosophy of Science* 85 (5):1078-1089.

Robins S. K. (2020). Stable Engrams and Neural Dynamics. *Philosophy of Science*. 87(5):1130-1139. https://doi.org/10.1086/710624

Robins, S. K. (2023). The 21st century engram. *WIREs Cognitive Science,* e1653. https:// doi.org/10.1002/wcs.1653

Roediger, H.L. (2007) Encoding. In H.L. Roediger, Y. Dudai & S.M. Fitzpatrick (eds.) *Science of Memory: Concepts* (pp. 17-21). Oxford University Press.

Root, C. M., Denny, C. A., Hen, R., & Axel, R. (2014). The participation of cortical amygdala in innate, odour-driven behaviour. *Nature*, *515*(7526), 269–273. https://doi.org/10.1038/nature13897

Ross, L.N. (2021). Causal Concepts in Biology: How Pathways Differ from Mechanisms and Why It Matters. *British Journal for the Philosophy of Science* 72 (1):131-158.

Ross, L.N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese* 202 (2):1-19 https://doi.org/10.1007/s11229-023-04281-5

Rota, G. C. (1985). The barrier of meaning. Lett Math Phys 10, 97–99. https://doi.org/10.1007/BF00398144

Roy, D. S., Arons, A., Mitchell, T. I., Pignatelli, M., Ryan, T. J., & Tonegawa, S. (2016). Memory retrieval by activating engram cells in mouse models of early Alzheimer's disease. *Nature*, *531*(7595), 508–512. https://doi.org/10.1038/nature17172

Roy, D. S., Park, Y. G., Kim, M. E., Zhang, Y., Ogawa, S. K., DiNapoli, N., Gu, X., Cho, J. H., Choi, H., Kamentsky, L., Martin, J., Mosto, O., Aida, T., Chung, K., & Tonegawa, S. (2022). Brain-wide mapping reveals that engrams for a single

memory are distributed across multiple brain regions. *Nature communications*, *13*(1), 1799. https://doi.org/10.1038/s41467-022-29384-4

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current opinion in neurobiology*, *58*, 141–147. https://doi.org/10.1016/j.conb.2019.08.005

Rule, M. E., & O'Leary, T. (2022). Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(7), e2106692119. https://doi.org/10.1073/pnas.2106692119

Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A., & Tonegawa, S. (2015). Memory. Engram cells retain memory under retrograde amnesia. *Science (New York, N.Y.)*, *348*(6238), 1007–1013. https://doi.org/10.1126/science.aaa5542

Ryan, T. J., Ortega-de San Luis, C., Pezzoli, M., & Sen, S. (2021). Engram cell connectivity: an evolving substrate for information storage. *Current opinion in neurobiology*, *67*, 215–225. https://doi.org/10.1016/j.conb.2021.01.006

Ryan, T. J., & Frankland, P. W. (2022). Forgetting as a form of adaptive engram cell plasticity. *Nature reviews. Neuroscience*, *23*(3), 173–186. https://doi.org/10.1038/s41583-021-00548-3

Ryle, G. (1949). *The Concept of Mind: 60th Anniversary Edition*. Hutchinson & Co.

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive psychology*, *139*, 101527. https://doi.org/10.1016/j.cogpsych.2022.101527

Sacktor T. C. (2011). How does PKMζ maintain long-term memory?. *Nature reviews. Neuroscience*, *12*(1), 9–15. https://doi.org/10.1038/nrn2949

Saigusa, T. *et al.* (2008) 'Amoebae Anticipate Periodic Events', *Physical Review Letters*. American Physical Society, 100(1), p. 18101. doi: 10.1103/PhysRevLett.100.018101.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question?. *Nature reviews. Neuroscience*, *22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. (2001). Forgotten ideas, neglected pioneers: Richard Semon and the story of memory. Psychology Press

Schulte, P. & Neander, K. (2022). Teleological Theories of Mental Content, in E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Summer 2022 Edition)*, https://plato.stanford.edu/archives/sum2022/entries/content-teleological/

Semon, R. (1921). *The Mneme*. George Allen & Unwin. (Original work published 1904)

Semon, R. (1923). *Mnemic Psychology*. (B. Duffy, trans.) George Allen & Unwin. (Original work published 1909)

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Sherry, D.F., & Hoshooley, J.S. (2010). Seasonal hippocampal plasticity in food-storing birds. Philosophical Transactions of the Royal Society B: Biological Sciences, 365, 933 - 943.

Shpokayte, M., McKissick, O., Guan, X., Yuan, B., Rahsepar, B., Fernandez, F. R., Ruesch, E., Grella, S. L., White, J. A., Liu, X. S., & Ramirez, S. (2022). Hippocampal cells segregate positive and negative engrams. *Communications biology*, *5*(1), 1009. https://doi.org/10.1038/s42003-022-03906-8

Simon, H. A. (2002). Near decomposability and the speed of evolution. *Industrial and Corporate Change, 11*(3), 587-600.

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor perspectives in biology*, *7*(8), a021766. https://doi.org/10.1101/cshperspect.a021766

Stanley, K.O., & Lehman, J. (2015). *Why Greatness Cannot Be Planned*. Springer International.

Steadman, P. E., Xia, F., Ahmed, M., Mocle, A. J., Penning, A. R. A., Geraghty, A. C., Steenland, H. W., Monje, M., Josselyn, S. A., & Frankland, P. W. (2020). Disruption of Oligodendrogenesis Impairs Memory Consolidation in Adult Mice. *Neuron*, *105*(1), 150–164.e6. https://doi.org/10.1016/j.neuron.2019.10.013

Sterling, L. & Laughlin, S. (2015). *Principles of Neural Design*. MIT Press.

Stone, J.V. (2022). *Information Theory: A Tutorial Introduction*. (2nd ed). Sebtel Press.

Sun, X., Bernstein, M. J., Meng, M., Rao, S., Sørensen, A. T., Yao, L., Zhang, X., Anikeeva, P. O., & Lin, Y. (2020). Functionally Distinct Neuronal Ensembles within the Memory Engram. *Cell*, *181*(2), 410–423.e17. https://doi.org/10.1016/j.cell.2020.02.055

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems.

Nature neuroscience, *26*(8), 1438–1448. https://doi.org/10.1038/s41593-023-01382-9

Tanaka, K. Z., Pevzner, A., Hamidi, A. B., Nakazawa, Y., Graham, J., & Wiltgen, B. J. (2014). Cortical representations are reinstated by the hippocampus during memory retrieval. *Neuron*, *84*(2), 347–354. https://doi.org/10.1016/j.neuron.2014.09.037

Terranova, J. I., Yokose, J., Osanai, H., Ogawa, S. K., & Kitamura, T. (2023). Systems consolidation induces multiple memory engrams for a flexible recall strategy in observational fear memory in male mice. *Nature communications*, *14*(1), 3976. https://doi.org/10.1038/s41467-023-39718-5

Tinbergen, N. (1963). "On aims and methods of ethology." Zeitschrift für Tierpsychologie 20:410-433.

Tomé, D. F., Zhang, Y., Aida, T., Mosto, O., Lu, Y., Chen, M., Sadeh, S., Roy, D. S., & Clopath, C. (2024). Dynamic and selective engrams emerge with memory consolidation. *Nature neuroscience*, *27*(3), 561–572. https://doi.org/10.1038/s41593-023-01551-w

Tonegawa, S., Liu, X., Ramirez, S., & Redondo, R. (2015a). Memory Engram Cells Have Come of Age. *Neuron*, *87*(5), 918–931. https://doi.org/10.1016/j.neuron.2015.08.002

Tonegawa, S., Pignatelli, M., Roy, D. S., & Ryan, T. J. (2015b). Memory engram storage and retrieval. *Current opinion in neurobiology*, *35*, 101–109. https://doi.org/10.1016/j.conb.2015.07.009

Tsien R. Y. (2013). Very long-term memories may be stored in the pattern of holes in the perineuronal net. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(30), 12456–12461. https://doi.org/10.1073/pnas.1310158110

Turing, A. M. (1936), “On Computable Numbers, with an Application to the Entscheidungsproblem,” *Proceeding of the London Mathematical Society*, 42(1): 230–265

Vardalaki, D., Chung, K., & Harnett, M. T. (2022). Filopodia are a structural substrate for silent synapses in adult neocortex. *Nature*, *612*(7939), 323–327. https://doi.org/10.1038/s41586-022-05483-6

Varela, F. G., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Currents in modern biology*, *5*(4), 187–196. https://doi.org/10.1016/0303-2647(74)90031-8

Varela F. J., Thompson E., & Rosch E. (1991). *The Embodied Mind*. MIT Press.

Vetere, G., Kenney, J. W., Tran, L. M., Xia, F., Steadman, P. E., Parkinson, J., Josselyn, S. A., & Frankland, P. W. (2017). Chemogenetic Interrogation of a Brain-wide Fear Memory Network in Mice. *Neuron*, *94*(2), 363–374.e4. https://doi.org/10.1016/j.neuron.2017.03.037

von Neumann, J. (1945). First draft of a report on the EDVAC. Republished *IEEE Annals of the History of Computing*, vol. 15, no. 4, pp. 27-75, 1993, https://doi.org/10.1109/85.238389

von Neumann, J. (1958) *The Computer and the Brain*. New Haven: Yale University Press

Wang, C., Yue, H., Hu, Z., Shen, Y., Ma, J., Li, J., Wang, X. D., Wang, L., Sun, B., Shi, P., Wang, L., & Gu, Y. (2020). Microglia mediate forgetting via complement-dependent synaptic elimination. *Science (New York, N.Y.)*, *367*(6478), 688–694. https://doi.org/10.1126/science.aaz2288

Wang, L., Gillis-Smith, S., Peng, Y., Zhang, J., Chen, X., Salzman, C. D., Ryba, N. J. P., & Zuker, C. S. (2018). The coding of valence and identity in the mammalian taste system. *Nature*, *558*(7708), 127–131. https://doi.org/10.1038/s41586-018-0165-4

Wilf, M., Strappini, F., Golan, T., Hahamy, A., Harel, M., & Malach, R. (2017). Spontaneously Emerging Patterns in Human Visual Cortex Reflect Responses to Naturalistic Sensory Stimuli. *Cerebral cortex (New York, N.Y. : 1991)*, *27*(1), 750–763. https://doi.org/10.1093/cercor/bhv275

Wilbanks, E. G., Doré, H., Ashby, M. H., Heiner, C., Roberts, R. J., & Eisen, J. A. (2022). Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *The ISME journal*, *16*(8), 1921–1931. https://doi.org/10.1038/s41396-022-01242-7

Wiltgen, B. J., & Tanaka, K. Z. (2013). Systems consolidation and the content of memory. *Neurobiology of learning and memory*, *106*, 365–371. https://doi.org/10.1016/j.nlm.2013.06.001

Wit, C. B., & Hiesinger, P. R. (2023). Neuronal filopodia: From stochastic dynamics to robustness of brain morphogenesis. *Seminars in cell & developmental biology*, *133*, 10–19. https://doi.org/10.1016/j.semcdb.2022.03.038

Wheeler, A. L., Teixeira, C. M., Wang, A. H., Xiong, X., Kovacevic, N., Lerch, J. P., McIntosh, A. R., Parkinson, J., & Frankland, P. W. (2013). Identification of a functional connectome for long-term fear memory in mice. *PLoS computational biology*, *9*(1), e1002853. https://doi.org/10.1371/journal.pcbi.1002853

y Arcas, B. A. (2023). AI and the Barrier of Meaning 2, Santa Fe Institute Conference, April 24—26. Available at https://youtu.be/uwUgYLaWTXc?si=m23b770AukT7ld7Q

Zador A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, *10*(1), 3770. https://doi.org/10.1038/s41467-019-11786-6

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Körding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Olshausen, B., Pouget, A., … Tsao, D. (2023). Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nature communications*, *14*(1), 1597. https://doi.org/10.1038/s41467-023-37180-x