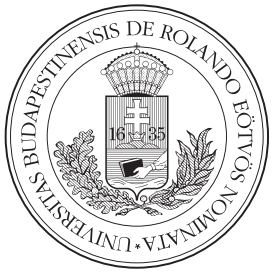
Eötvös Loránd University

Institute of Physics

School of Statistical Physics, Biological Physics
and Physics of Quantum Systems

Towards a normative account of human memory

Doctoral dissertation



Supervisor:
Gergő Orbán, Ph.D.
Senior Research Fellow
Wigner RCP

Author:
David G. Nagy
Junior Research Fellow
Wigner RCP

Head of Doctoral School:
Jenő Gubicza, D.Sc.

Head of Doctoral Program:
Gábor Horváth, D.Sc.

Budapest, 2023
DOI: 10.15476/ELTE.2023.028

Acknowledgements

First, I would like to express my gratitude to my advisor, Gergő Orbán, for his invaluable guidance, support, and encouragement throughout my studies as well as for providing me with the freedom to explore my interests. I am also deeply grateful for his establishment and tireless maintenance of an outstanding computational neuroscience lab in Budapest, where I have met many wonderful people.

I would like to thank my former advisor Szabolcs Káli for introducing me to computational neuroscience as well as co-supervising the initial project that turned into my doctoral research.

Among lab members, I would like to extend a special thanks to Mihály Bányai for all the thought-provoking discussions and debates, and for being a constant source of support and mentorship. I am also fortunate that Balázs Török joined the lab as a PhD student soon after I did, it was great to collaborate with him and I am grateful for his friendship throughout the years. I would also like to thank all current and former members of the lab, including Csenge Fráter, Merse E. Gáspár, Márton Hajnal, Ferenc Csikor, Balázs Meszéna, and Anna Székely for their contributions and camaraderie. In addition to the lab, I would like to acknowledge the wider computational neuroscience community in Budapest, including Zoltán Somogyvári at Wigner, Balázs Ujfalussy at KOKI and the cognitive science community at CEU, including Máté Lengyel, József Fiser and their students, especially Oana Stanciu and Ádám Koblinger.

I am grateful to Chris Summerfield for giving me the opportunity to join his lab for an internship and collaborate with Timo Flesch on an exciting internship project. I would also like to thank his other lab members and especially Leonie Glitz and Ron Dekker for being so welcoming and making Oxford feel like home.

During my university studies, I am grateful to ELTE for providing the opportunity to take classes from across the university, allowing me to explore a broad range of interests. I feel especially fortunate to have found the courses of László E. Szabó, whose ideas profoundly influenced my thinking. I am also thankful for having had the chance to learn from the insightful perspectives of István Csabai, Antal Jakovác, and Gyula Dávid, and for having had such wonderful classmates, especially Balázs, Lilla, and Sándor.

I am deeply thankful to Eszter, Vera, Gergő, Gábor, Balázs, Ádám, Zoli and Ákos.

I would like to extend a special thanks to Csenge.

Finally, I am immensely grateful to my family: Csaba, Nóra, Csenge and Botond as well as Kata, Anna and Sári.

Contents

Chapter 1

Introduction

The human brain is constantly bombarded with sensory information, as photons, vibrations, molecules, and other messengers of the state of affairs in the outside world arrive at the senses, being then translated into the activities of sensory neurons. From this constant barrage of sense data, what should the human brain retain and what should it forget? That is the central question driving this thesis. Answering the question could shed light on how human memory works and why it works that way, constituting a key piece in our understanding of the brain. Furthermore, it could result in uncovering principles helpful for building artificial agents with human-like capabilities. However, as we will see, this question is intertwined with a host of others, such as what representations does the brain transform this sensory input into, what concepts does it use to interpret it and how does it construct models out of them?

In this thesis I take a normative approach, working towards an account of how the brain works starting from an analysis of the computational problems that it has evolved to solve. A primary objective of normative approaches is to develop vocabulary and formal tools for thinking about these problems, which will enable a characterisation of the space of optimal and near-optimal solutions. To the extent that evolution shaped the brain to solve the identified problems, the optimal solutions can serve as useful points of comparison and both explain and predict the brain's behaviour in a parsimonious way. Due to the focus on optimal solutions, the normative approach has a large overlap in interests with machine learning and hopefully, both the vocabulary and the solutions can be exploited for the construction and improvement of artificial

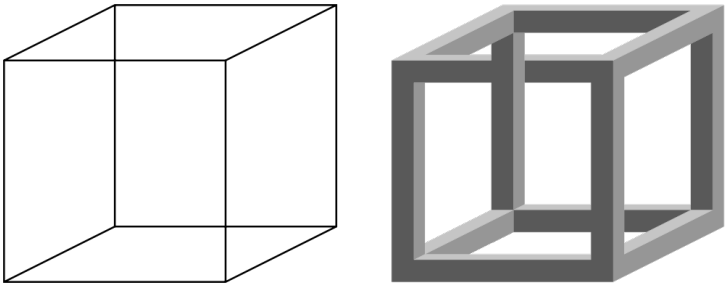


Figure 1.1: **Left,** Necker cube. The cube can be perceived either as protruding towards the bottom left, or the top right, depending on which square-shaped component is assumed to be closer to the viewer. **Right,** a visual illusion called the impossible cube, playing on the ambiguity present in the Necker cube. The original version of this illusion appears in Escher's 1958 lithograph *Belvedere*. Cognitive illusions such as the impossible cube can often be explained at the computational level without making commitments to the 'hardware' implementation. For detailed examples see section 1.2.2.

agents.

My approach in this work aligns with the rational analysis framework of Anderson in the field of cognitive science [1] and the top-down analysis of Marr and Poggio's levels of analysis for neuroscience [2]. Marr famously argued that complex information processing systems such as the brain should span multiple complementary levels of analysis, specifically the computational, the algorithmic and the implementation levels. In a normative analysis, we start from the top down, asking first the questions of what computational problem needs to be solved by the system and what algorithmic procedures can possibly solve it. Once we have mapped the space of possible solutions on this 'software' level, we can ask how these algorithms can be implemented in the physical 'hardware' of the brain. A bottom-up analysis would proceed in the opposite direction, starting with building blocks such as detailed models of neurons and attempting to understand how they can be combined to produce behaviours similar to the human brain [3, 4].

In some cases, the origin of observed phenomena can be clearly identified with the consequences of theories at a specific level of analysis. For example, the bistable perception of a Necker cube originates in the existence of multiple possible interpretations of a projection of a 3D object to a 2D screen – a computational level explanation that does not require many biological commitments to how the visual cortex works (Fig. 1.1). However, it is important

to note that these levels of description do not always separate clearly, and starting our analysis from the top down is a trade-off that carries risks with it. For instance, the extent to which the computational problems being studied exerted evolutionary pressure on the development of the brain can be overestimated. A second issue is that there are multitudes of constraints on many different scales that the brain has to satisfy and they can interact in non-trivial ways across these levels of analysis.

In this thesis I follow a top down approach. Therefore, in the rest of the chapter, I will first attempt to identify the computational problems that memory has to solve and introduce some of the formal ideas and frameworks that previous research has converged on. I introduce the concept of memory systems, most importantly the distinction between episodic and semantic memory. In Chapter 2, I identify a fundamental computational challenge for learning and propose that it provides a normative rationale for the existence of an episodic memory system which retains seemingly irrelevant details as part of the memory trace. Then, in Chapter 3, we explore the idea that semantic memory supports the episodic memory system by enabling efficient storage of episodic memories through offering a model for a lossy compression of experiences. I demonstrate that this this process offers a parsimonious unifying explanation of a variety of memory distortions that previous research have identified. Finally, in Chapter 4 I explore two examples of how some of the computational and algorithmic level ideas discussed in the previous chapters can be implemented in neural networks.

1.1 The computational problem of memory

An organism's survival critically depends on its ability to gather information from the environment through its senses and to affect it through actions. A straightforward approach to choosing actions is to base them solely on current sensory input, for example moving in the direction of a chemical gradient to locate higher concentrations of nutrients. However, a powerful strategy developed by complex organisms is to anticipate changes in the environment before they happen, enabling the organism to act earlier than competitors or before the opportunity to affect the unfolding of environmental processes is gone.

Figure 1.2: Internal representation of the environment in the human brain. Drawing made with Midjourney in the style of Escher.

This ability to predict future sensory input confers a tremendous advantage to complex organisms and has been proposed as the key driving force behind the evolution of the neocortex [5]. Therefore, my starting point in this thesis is that supporting the capacity for prediction is one of the main goals of memory in complex organisms.

How is prediction possible? To anticipate changes in the environment, the organism needs to reshape part of its body[1] into an artefact that in some respect mirrors the evolution of external processes. A general strategy is to create a simplified internal copy of the environment, an *internal model* or *internal representation* that can be used to simulate the original process and observe the outcomes[2]. These simulations can then be used to evaluate the consequences of actions the organism might consider taking.

[1] Or part of its environment.

[2] This is not to say that the brain has to maintain a single consistent internal model. There could be various competing models for different environmental contexts. Nor does this mean that all actions have to be based on simulations from explicit models, e.g. in situations that require rapid reactions a more direct mapping between stimulus and action might be preferred.

Internal models, effective theories

Internal models are limited by the physical size of the brain as well as metabolic constraints on computation, and available resources are severely mismatched to the complexity of maintaining a faithful simulation. Crucially, a model is only useful for prediction if the speed of simulation can exceed the speed at which environmental processes unfold. Therefore, the brain needs to find computational shortcuts by constructing a simplified representation that only includes environmental variables that are most relevant for the decisions it has to make. While to my knowledge there is no widely accepted and generally applicable normative framework for how such simplified representations should be constructed, specific cases of the problem have been solved in various fields. In physics for example, such simplified models that describe a complex system on a macroscopic scale are known as *effective theories*. Similar concepts have also been formalised as reduced models [6] or abstractions [7]. Intuitively, an effective theory is a simplified or coarse-grained macroscopic description of a system that emerges from a low-level microscopic description by averaging over or leaving out interactions, objects and other complexities that are inconsequential to the macroscopic description. The macroscopic description often uses qualitatively different concepts from the microscopic one, which is often termed *emergence*. One of the simplest examples originates in statistical physics, describing how macroscopic quantities such as temperature and magnetisation emerge from a high-level description of a large grid of identical nuclear spins. However, the human brain is typically faced with environmental contexts that are far more complex and rarely tractable analytically. I would argue that for many of these contexts the best effective models are constituted by board games and video games [8], which can be seen as simplified descriptions of environmental contexts, where objects and interactions are chosen such that they mimic real world systems but stay closer to the representational capacity of the brain (Fig. 1.3).

Semantic and episodic memories

Returning to our original question of what to retain from a continuous stream of sensory experience, we conclude that it is an important goal to

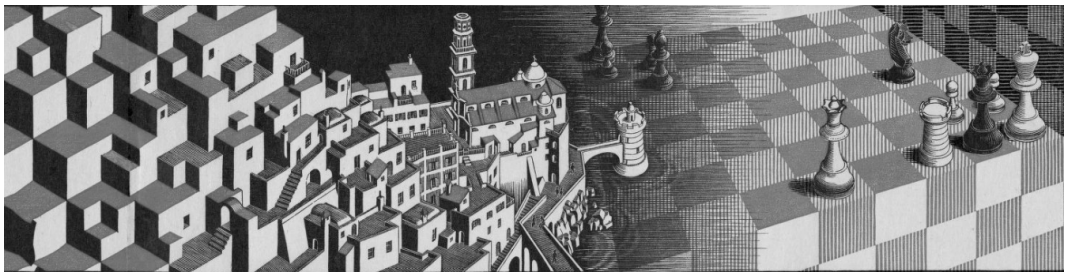


Figure 1.3: Excerpt from Escher, *Metamorphosis III*. Here, it is used as an illustration of the process of constructing effective theories or abstractions: relevant concepts transform and new ones emerge as we look at the same system at successively larger scales, eliminating aspects deemed irrelevant at each step. A set of elementary building blocks viewed at a larger scale might be better described as a city, and the operation of a city at a different scale and from the viewpoint of certain agents might be best captured in the form of a board game such as chess.

retain information that is necessary for supporting the organisms ability to predict. In particular, we identified a key piece of this challenge is to extract the information needed to construct and update an internal model, a simplified representation of the environment. Indeed, formalising this information selection problem as maximising predictive ability while retaining as little total information as possible [6], it can be shown that the optimal solution is to learn the generative model that produces the observations [9]. Furthermore, the amount of information that the agent has regarding future observations is precisely the information that its approximate generative model contains regarding the true process.

In the study of human memory, maintaining general knowledge about how the world works has been identified as one of the two fundamental memory systems that comprise long-term memory. This first system is called *semantic memory*, responsible for storing common facts about the world (such as how common household objects work for example), as well as the meanings of words and concepts. In formal treatments, semantic memory is often defined as a generative model of the environment [10, 11, 12]. The second system, known as *episodic memory*, is responsible for preserving specific events and experiences that have occurred in a person's life. Episodic memories can be likened to vivid snapshots of lived experience and are believed to be closely related to the ability for 'mental time travel' in humans.

In this thesis, I aim towards a normative account of the semantic and

episodic memory systems and their interactions. Both memory systems retain information from past experiences but in complementary ways and I aim to uncover their normative role in supporting the goals of biological agents. In the following section, I focus on semantic memory and discuss the normative framework of probabilistic generative models in which its main function of representing knowledge that supports prediction and other cognitive functions can be formalised. A proposal for the normative rationale behind the existence of episodic memory is one of the main contributions of this thesis and will be the subject of Chapter 2.

1.2 Memory as knowledge: semantic memory

Maintaining knowledge of the environment is a crucial responsibility of human memory, and this responsibility rests primarily with the semantic memory system. This section aims to provide an overview of key ideas from prior research that can serve as a foundation for a normative account of how this responsibility can be met. First, I will outline the reasons for why knowledge representation in the brain, particularly in semantic memory, should be in the form of a probabilistic generative model. Then, I briefly introduce the theoretical foundations of probabilistic models, including graphical models and probabilistic programs. I show how this approach addresses many of the key challenges in cognition, including perception as unconscious inference. Finally, we examine how learning, or the construction of a representation of the environment over an organism's lifetime, can be achieved within the framework of probabilistic inference.

1.2.1 Knowledge representation

How can knowledge be represented in a physical system such as the brain? I have argued that to perform predictions, complex organisms shape part of their brain into an artefact that in some sense mirrors the environment. Human-made prediction devices, for example, orreries such as the Antikythera mechanism serve as intuitive examples of this notion of mirroring. The Antikythera mechanism is an ancient device from the 1st or 2nd century BCE and is con-

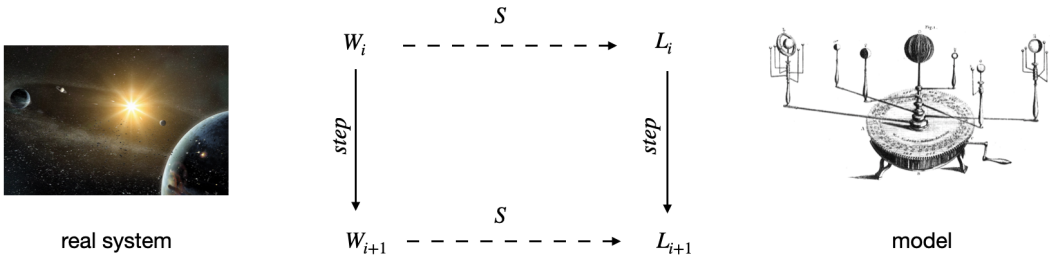


Figure 1.4: Representing knowledge of the environment in physical systems. The orrery on the right serves as a model of the solar system. The states of the real system, W_i, correspond to representational states of the orrery, L_i, and this correspondence, S, is maintained over the time evolution of the systems, allowing the user to predict the relative locations of the celestial bodies in the future.

sidered one of the first computers. The device had multiple dials representing observable astronomical information including the positions of the Sun and Moon, the moon phase, eclipses, and possibly the locations of planets. By turning the hand crank on the side of the device, interlocking gears within the mechanism moved the dials to reflect a future state of celestial objects, allowing the operator to predict future events, such as the date of the next eclipse, by reading off the dials. The Antikythera device is a physical system, parts of which (dial positions) map onto important features of another physical system (celestial body locations), and crucially, these mappings are preserved under certain transformations (e.g. time evolution). While a precise definition of the concepts of representation and computation are difficult questions [13, 14], this example intuitively illustrates how a physical system can represent knowledge regarding an external physical system (Fig. 1.4).

Note that to be useful for prediction, the variables on the dials had to transform in the same way as the positions of the celestial bodies, even though the mechanism in the former case was a system of gears whereas in the latter it was the force of gravity between planets and stars. This raises the important question of what are the limits of such representation. Which kinds of mechanisms can model which other kinds of mechanisms? Perhaps most importantly from our point of view, which kinds of systems can be modelled by neurons? One of the most important discoveries of the 20th century is that surprisingly, all ways of defining effective computation led to the same set of functions. The definition that most closely aligns with computation with physical devices was developed

by Turing, using abstract machines that process discrete symbols arranged on an infinite tape using a set of defined instructions known as Turing machines. Inspired by his work, McCulloch and Pitts [15] analysed idealised neural networks and suggested that recurrent versions of these networks coupled with memory could calculate the same functions as Turing machines. Although this was just a conjecture [3], such a construction was developed much later by others [17]. Turing also demonstrated that it was possible to build a universal Turing machine (UTM) that could simulate any other Turing machine by encoding its instructions on the tape. These results led to the formulation of the Church-Turing thesis, which asserts that any intuitively computable function is computable by some TM and consequently by any other sufficiently intricate (Turing complete) mechanism. This notion of *computational universality*, the independence of the computation from the concrete physical mechanism is the basis for the existence of software separately from its hardware level implementation. Furthermore, it suggests that the brain being made of neurons can be consistent with it performing computations that are easier to understand in formal systems adapted to the high-level challenges that it faces, providing the basis for a top-down analysis.

Logic

The universality of computation suggests that we may begin our analysis in a framework ideally suited to the problems of representing knowledge in a format that can support predictions, simulations and reasoning. We start with logic, the system that describes the laws of sound reasoning that originated with Aristotle and was refined by many others [18, 19]. Specifically, we introduce a simple variant called propositional logic. To construct a model in propositional logic, we have to specify the distinct states of the system. We can decompose these states using state variables (atomic propositions). The full model can be encoded as a truth table, where columns represent the atomic variables in the system and the rows list all possible combinations of values for those variables. Each row in the truth table represents a state of the system, with the values for the atomic variables given in the corresponding columns.

[3] It is a common misconception that they have proved this, for details see Piccinini, 2020 [16].

Cough	Flu	TB	Possible
1	1	1	1
1	1	0	1
1	0	1	1
1	0	0	0
0	1	1	0
0	1	0	0
0	0	1	0
0	0	0	1

Cough	Flu	TB	Possible
~~1~~	~~1~~	~~1~~	~~1~~
~~1~~	~~1~~	~~0~~	~~1~~
1	0	1	1
~~1~~	~~0~~	~~0~~	~~0~~
~~0~~	~~1~~	~~1~~	~~0~~
~~0~~	~~1~~	~~0~~	~~0~~
~~0~~	~~0~~	~~1~~	~~0~~
~~0~~	~~0~~	~~0~~	~~1~~

Figure 1.5: Truth tables and inference in a toy diagnostic model. **Left,** the original truth table. **Right,** the truth table with lines incompatible with the query '*If the patient is coughing but doesn't have the flu, is it TB?*' crossed out.

The last column of the truth table defines whether the combination of truth values represented in the row is a possible state of the system (see Fig 1.5. for an example).

In order to be able to make inferences in a logical system represented by a truth table, we can simply cross out all states that do not match the query or are not possible. For example, if we make the query: '*If the patient is coughing but doesn't have the flu, is it TB?*', we need to cross out all states where $cough \neq 1$ or $flu \neq 0$ and all states where $possible \neq 1$ (Fig. 1.5). The only possible state left is row 3, where $TB = 1$ and therefore the answer to the question in this simple model is yes.

The first issue with this method becomes apparent if we start refining the model by adding new variables: the number of rows scales exponentially with the number of variables. This makes both representing the table and answering queries very cumbersome and therefore it is useful to introduce logical operators. Each operator is defined by a smaller table called a conditional truth table (Fig. 1.6). Using these CTTs to decompose the full truth table makes the definition of the model much shorter. For example, using logical operators, the truth table in Fig 1. can be described simply by:

$$(\text{flu} \vee \text{TB}) \leftrightarrow \text{cough}$$

Moreover, these operators enable more efficient means of performing inference than the method of crossing out lines in the truth table:

$$A \leftrightarrow B, A \vdash B$$

A	B	A $\leftrightarrow$ B
1	1	1
1	0	0
0	1	0
0	0	1

A	B	A $\vee$ B
1	1	1
1	0	1
0	1	1
0	0	0

Figure 1.6: Conditional truth tables for the IFF operator (A $\leftrightarrow$ B) and the OR operator (A $\vee$ B).

$$A \vee B, \neg A \vdash B,$$

where $\vdash$ stands for logical entailment. For example, to answer the question '*If the patient is coughing but doesn't have the flu, is it TB?*', or in logical notation:

$$\text{cough} \wedge \neg\text{flu} \vdash \text{TB},$$

we can use the above two inference rules by substituting into the first rule,

$$\text{cough} \leftrightarrow (\text{flu} \ \vee \ \text{TB}) \ , \text{cough} \vdash (\text{flu} \ \vee \ \text{TB}),$$

and then substituting the above result into the second inference rule, we get:

$$\text{flu} \vee \text{TB}, \neg\text{flu} \vdash \text{TB}.$$

Therefore, we can conclude that in this situation, TB is true.

Despite the introduction of these logical operators, the expressive power of propositional logic remains limited. However, let us consider a different kind of problem first. Assume the question was '*If the patient is coughing, does he have TB?*'. If we attempt to use the method of crossing out incompatible and impossible states of the world, we are left with only three possible states: two in which the patient has TB and one in which he does not. This means that there are queries which are undecidable in the system and in such cases, logic fails to provide a useful answer.

Undecidability and degrees of belief

Unfortunately in practical situations, the majority of queries that the agent is concerned with will fall into the category of undecidable or uncertain. However, there is a simple modification we can make to allow the system to deal with uncertain knowledge: we can allow the 'possible' column to take

on non-binary values ranging from zero to one, corresponding to the agent's degree of belief or plausibility that the world is in that state. How should the agent choose these numbers? Two well-known arguments exist regarding how degrees of belief should be consistently assigned to states of the world.

First, Cox [20] attempted to demonstrate that if an agent's degrees of belief are expressed as real numbers and satisfy axioms encoding common sense requirements, such as the requirement that different ways of arriving at an answer from the same information should lead to the same outcome, then these beliefs must adhere to the axioms of probability theory. Therefore, in extending logic to handle uncertainty, the degrees of belief should be chosen such that they satisfy the rules of classical probability theory. Cox's theorem can be seen alternative axiomatisation of probability theory, equivalent to Kolmogorov's axioms but with a different motivation. In this correspondence with probability theory, queries are formalised as the calculation of conditional probabilities, where the condition defines the query. For example, our query in the diagnosis system would correspond to the computation of $P(\text{TB}|\text{cough}, \neg\text{flu})$. It can be shown that the computation of conditional probabilities is equivalent to the method of crossing out the incompatible rows of the truth table and renormalising the remaining rows to sum to unity.

The second argument, called the *Dutch book argument* [21], assumes that the agent is willing to make bets according to his degrees of belief. The argument shows that if these degrees of belief fail to follow certain consistency rules, there will exist a set of bets that the agent would accept as fair, even though it would guarantee a loss. Such a set of bets is called a 'dutch book' by professional bookmakers. As in Cox's theorem, the consistency rules that degrees of belief have to satisfy correspond to the rules of classical probability theory.

Taken together, these two arguments are key pieces in motivating the use of probability theory as a normative framework for understanding human cognition. In particular, they suggest that the agent's internal models should be formalised as probabilistic models of the environment with the probabilities corresponding to how plausible the agent considers each state to be. The agent can query its representation through computing conditional probabilities, which corresponds to probabilistic inference. I will discuss in more detail

how various cognitive functions can be cast as probabilistic inference in section 1.2.2, but first, we return to the problem of expressivity.

Graphical models

Joint probability tables, the probabilistic extensions of truth tables suffer from the same '*curse of dimensionality*' as their counterparts from logic, that is, the problem of exponential scaling with the number of variables. Similarly to logic, there are extensions of probability theory that exploit compositionality to increase the expressivity of probability tables. Analogously to the way that truth tables can be decomposed using operators defined by conditional truth tables, joint probability tables can be decomposed using smaller conditional probability tables. This trick leads to the concept of directed graphical models in probability theory. In graphical models, each variable in the model corresponds to a node in a directed acyclic graph (DAG), where the parents of each node are the variables which are included in the conditional probability table for the variable represented by the node. In addition to allowing for a more compact representation of the joint probability table, graphical models also enable inference algorithms that make certain kinds of queries such as statistical dependencies between the variables more efficient to compute.

Since graphical models play a central role in probabilistic models of cognition, we consider them in more detail. In order to see how graphical models enable a more compact representation of the joint distribution, we first use the chain rule for probabilities to factorise it for an arbitrary ordering $X_1, X_2, .., X_n$ of n random variables as

$$P(x_1, x_2, \ldots, x_n) = P(x_1|x_2, \ldots, x_n)P(x_2|x_3, \ldots, x_n) \ldots P(x_n).$$

If the conditional probability of a certain variable X_i is only dependent on a subset of the variables that it is conditioned on, and we call this subset pa_i (the Markovian parents of X_i) then we have

$$P(x_i|x_{i+1}, ..., x_n) = P(x_i|pa_i).$$

Figure 1.7: An illustration of the explaining away effect for probabilistic models. **Left:** We suppose that burglaries and earthquakes are unrelated and rare, but both cause the house alarm to signal. It follows that being informed that the alarm went off increases our concern about both an earthquake and a burglary being in the house. However, if, in addition to hearing the alarm, we also feel the ground trembling, it makes us less likely to think that we have also been the victims of burglary. **Right:** Direct observation of latents is not necessary for them to become dependent - if our model also includes the fact that valuables go missing after a break-in, then noticing both the alarm going off and the disappearance of possessions is enough to infer the presence of the thief and conclude that there was probably no earthquake.

This means that the joint distribution can be written as

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_I|pa_i),$$

which can greatly reduce the information needed to specify it (that would otherwise require $2^n - 1$ entries even for binary variables) by decomposing it into smaller distributions. Furthermore, we can construct a DAG that satisfies the same child-parent relationships with each variable having a corresponding node and each conditional dependence a directed edge. We call such a graph G a graph representation of distribution P. We introduce the following notation for independence between variables

$$X \perp Y \quad \Leftrightarrow \quad P(x) = P(x|y) \quad \Leftrightarrow \quad P(x,y) = P(x)P(y),$$

and for conditional dependence

$$X \perp Y|Z \quad \Leftrightarrow \quad P(x|z) = P(x|y,z) \quad \Leftrightarrow \quad P(x,y|z) = P(x|z)P(y|z).$$

The graph representation provides an economical encoding of the independence relations or correlation structure of the distribution. It also enables these relations to be read out via graph-search algorithms instead of algebraic

methods, and can be given a causal interpretation. We say that probabilistic influence can flow from variable X to variable Y through a set of variables Z if $X \not\perp Y|Z$. Independencies can be read off the graph representation by checking whether there exists a trail (a path where the edges can point in either direction) through which influence can flow between the variables in question [22]. The flow of influence is blocked by conditioning on nodes Z if an only if:

1. the trail contains $u \leftarrow m \leftarrow v$, such that $m \in Z$,
2. the trail contains $u \leftarrow m \rightarrow v$, such that $m \in Z$,
3. or the trail contains $u \rightarrow m \leftarrow v$ such that $m \notin Z$ and no descendants of m are in Z,

where u, m and v are nodes in graph G. An interesting dependence relation is made evident by these criteria, called selection bias in the statistical and *explaining away* in the AI literature. The third criteria implies that otherwise independent parents can become correlated if we condition on their mutual child. Informally, if we observe something that can be a cause of some other observation, then this reduces the need for an alternative explanation. This also happens if we don't observe the hidden cause, but infer it from some auxiliary observation. An illustration of this effect can be found in Fig 1.7.

While both the introduction of operators and inference rules in propositional logic and graphical models in probability theory are important conceptual steps in dealing with the curse of dimensionality of truth tables and probability tables respectively, their expressive power is still limited. Logic has been extended into higher order logics, for example through the introduction of predicates and quantifiers results in first order logic. An important development was λ-calculus, which in addition to corresponding to a higher-order logic, has equivalent expressive power to universal Turing machines and thereby capable of expressing any effectively computable function. λ-calculus can be directly extended to handle probabilities with the introduction of a random sampling operator. This extension, called stochastic λ-calculus, provides the basis for a Turing-complete formalisation of probabilistic models called probabilistic programs [23].

Probabilistic programs

Probabilistic programs provide a Turing-complete modelling language that allows for the construction of mental simulations of any computable generative process. Each program represents a probability distribution through the relative frequencies of its outputs when executed. Running the program simulates the environment and results in a possible world state that is consistent with the model's initial conditions. Each of these output states can be considered a sample from the distribution represented by the program and as the number of samples grows, the relative frequencies of the outputs tend towards their probabilities, providing a sampling or Monte Carlo representation (for more details on sampling representations see section 1.3.1).

Similarly to directed graphical models, an important advantage of probabilistic programs that they can represent causal models. They have been proposed as a framework for formalising core human competencies such as intuitive physics and intuitive psychology by enabling the rich mental simulations of causal processes that these require [24]. For example, intuitive physics has been proposed to be similar to physics engines in modern video games that are used to create environments for the players to interact with. These physics engines contain simplified objects with properties and approximate dynamics for updating the world state and usually also incorporate a graphics engine that renders the 3D visual scene from the perspective of the player. This hypothesis has been successfully applied in a range of human experiments [25, 8]. In relation to this thesis, probabilistic programs are important primarily as a general formalisation of the compositionally defined, open-ended hypothesis spaces that the human brain has to be able to navigate during learning. We explore this view of learning as a process of finding the probabilistic program that generates the observed data in section 1.2.3.

1.2.2 Probabilistic models of cognition

As discussed in the previous section, making inferences is a key part of knowledge representation. Probabilistic inference has emerged as a unifying language for modelling inference problems that arise in the fields of cognitive science, machine learning and computational neuroscience. In this section, I

provide an overview of how important cognitive functions, such as decision-making and perception, can be framed as probabilistic inference, supported by the world model stored in semantic memory.

Perception as inference

A fundamental challenge for the brain is the difference between the quantities it can directly measure through sensory neurons and those that are relevant to plans and decisions. Quantities in the former category are things like impacts of photons, vibrations in surrounding air, temperature and the presence of certain molecules, whereas the brain is more concerned with determining the presence and properties of objects, animals, individuals, and their thoughts and intentions. This process of inferring the latent quantities based on the directly measurable ones was termed unconscious inference by Helmholtz [26], alluding to the observation that introspectively we only have access to the result of this computation (the *percept*), not the raw sensory data (the *sensation*). This unconscious inference in perception can be cast as probabilistic inference in generative models. We have discussed that in the probabilistic framework, inference can be performed by computing conditional distributions. For a query, where we want to determine the latent or hidden variables given the observed value o', we can apply the definition of conditional distributions and apply the chain rule to the numerator to get

$$P(\text{hidden} \,|\, \text{observed} = o') = \frac{P(\text{observed} = o' \,|\, \text{hidden})\, P(\text{hidden})}{P(\text{observed} = o')}. \tag{1.1}$$

This expression for computing conditional distributions is called Bayes' theorem, and the method of computing the distribution of hidden variables given observations is termed Bayesian inference in statistics. In Bayes' theorem, each term has a specific name based on its role in updating beliefs as new evidence is observed. The distribution of the hidden variable before conditioning $p(\mathit{hidden})$ is called the *prior* distribution, reflecting that it represents the beliefs of the agent prior to the new experience. The prior is multiplied by the *likelihood* term which encapsulates the effect that the evidence has on the belief over

likely values for the unknown variable [4]. The normalised product of these two terms gives the updated beliefs over the hidden variable, therefore it is called the *posterior* distribution. The normalisation constant is called the *marginal likelihood* referring to the fact that it can be computed by marginalising over the hidden variable in the numerator. Eq. 1.1. can then be written as

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal likelihood}} .$$

Observing the terms in Eq. 1.1, we may notice that the likelihood and posterior share the same variables, but with their roles reversed. If we consider the interpretation we have introduced as Helmholtz's perception as unconscious inference (or more commonly just *perception as inference*), where the goal is to find the true but hidden state of the environment (h) given raw sensory data (o), this means that the answer to the question of 'what is the state of the environment if I observed o?' is deeply connected to that of 'what would I observe if the state of the environment was h?'. Specifically, in order to compute the posterior $p(h|o)$, we have to invert the model that describes how the observations are generated $p(o|h)$. Therefore, the latter direction is sometimes called a *forward probability*, generative direction, predictive direction or simulator whereas the former is sometimes called an *inverse probability* or *model inversion* (see Fig. 1.8, left).

Computing the likelihood, or in other words, finding hidden states consistent with the observation can be sufficient to perform inverse inference in some problems, however this is not always the case. Observations are often ambiguous due the ubiquity of non-injective mappings between physical quantities due to resource and other biological constraints on the sensory systems of agents, or inherent in the way the environment works. For example, the projection of three dimensional objects onto a two dimensional image on the retina is a non-injective process, resulting in many three dimensional objects having the same image and making the inverse mapping ambiguous (Fig. 1.8, right). In these cases, an expectation of likely states based on prior knowledge can help in finding the correct interpretation, which is the role of the prior distribution.

[4] Note that when we call this term the likelihood, it is understood that we mean it as a function of the hidden variable $p(\text{observed} \,|\, \cdot)$ and therefore it is an unnormalised function, not a probability distribution.

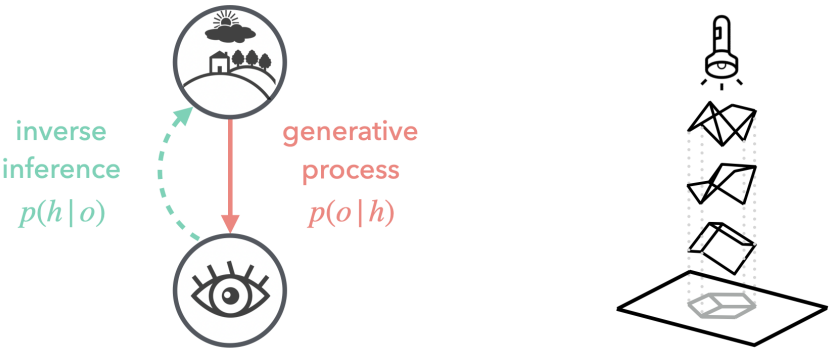


Figure 1.8: Perception as model inversion. **Left,** during perception, the generative model that describes how the latent factors in the environment generate the brain's sensory input is inverted using Bayesian inference in an unconscious process. **Right,** An example of the various sources of uncertainty in the unconscious inference process. A two dimensional shadow is consistent with a variety of three dimensional wire frames, and some additional source of knowledge, for example of environmental statistics, is necessary to find the most likely explanation. Shapes based on illustration in Kersten et al. [27].

Similar ambiguities are ubiquitous in perception both in vision, hearing and understanding language. In colour vision for instance, the directly observable input for the visual system is the distribution of photons falling on the retina (called the spectral power distribution, SPD), or more specifically its projection onto three basis functions, given by the response spectra of the S, M and L cone cells, defining a three dimensional colour space. However, this raw sensory input is not what is perceived as colour: for example, coal in the sunshine emits a lot more photons than snow during the night, however the former is perceived as black and the latter is perceived as white. As we have argued above, the goal of the visual system is no to present the raw sensory input, but to infer the latent causes that generate it. This involves inferring the material properties of surrounding objects and surfaces by estimating their reflectance, which is closer to what corresponds introspectively to the colour of an object. Reflectance combines with the light source SPD in a non-injective way, which is an additional source of ambiguity in colour vision. A striking cognitive illusion relying on non-injectivity can be seen on Fig. 1.9.

In language, the inference problem is to determine the speaker's intended meaning, based on the utterance. Language is overflowing with ambiguities both on the level of words and sentences. For example, consider the newspaper headline,

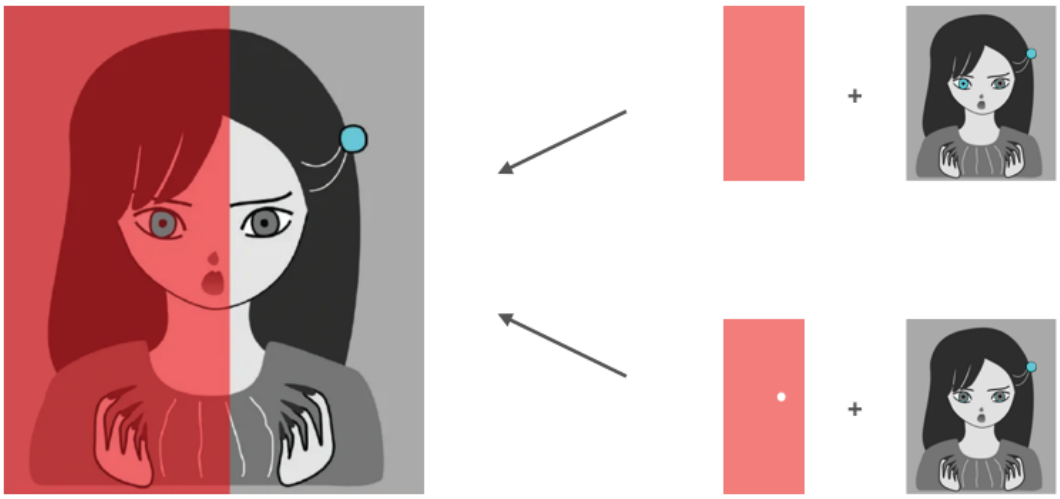


Figure 1.9: Visual illusion relying on perceptual ambiguity and the explaining away effect. **Left,** The spectral power distribution (SPD) arriving on our retina is the same for both eyes of the girl, yet they are typically perceived as different colours. The SPD arriving from the left side of the image strongly suggests a strong red light source (e.g. the girl could be standing in front of a projector) or alternatively a red filter, which explains away why the otherwise cyan coloured eye appears grey (see Fig. 1.7). **Right,** the same image is also consistent with another explanation: the entire left side of the image is illuminated by red light with the exception of the left eye, which is illuminated by normal light, revealing that it is the same gray colour as the other one. Seeing the left eye as cyan suggests that the unconscious inference process concludes that it would be an unlikely coincidence for the red illumination/filter to exclude the left eye specifically. The cyan-eyed girl illusion was created by prof. Akiyoshi Kitaoka.

'Drunk Gets Nine Months in Violin Case',

which is consistent with two grammatically correct interpretations of 'violin case', due to a tendency in natural language for the same word to have multiple possible meanings (polysemy). Furthermore, grammar itself is ambiguous, as exemplified by the following sentence [28]:

'Two cars were reported stolen by the Groveton police yesterday' .

The sentence can be parsed in multiple ways, associated with different meanings:

1. interpretation: 'Two cars were (reported stolen) by the Groveton police yesterday.'

2. interpretation: 'Two cars were reported (stolen by the Groveton police) yesterday.'

There is no difference in the likelihoods,

$$P(\text{sentence} \,|\, \text{meaning} = 1.) = P(\text{sentence} \,|\, \text{meaning} = 2.),$$

because either occurrence can be equally well communicated with the sentence. Encountering these sentences as actual headlines would however be unlikely to be noticeable for the reader[5], due to the unconscious inference process disambiguating the sentence through the use of prior expectations: it is uncommon for the police to steal cars whereas it is common for them to report thefts: $P(\text{meaning} = 1.) > P(\text{meaning} = 2.)$. By combining these pieces of knowledge in the posterior, we find that

$$P(\text{meaning} = 1. \,|\, \text{sentence}) > P(\text{meaning} = 2 \,|\, \text{sentence}).$$

Taken together, an important role of semantic memory in perception is to provide the generative model, which enables the process of unconscious inference converting sensory data into variables that are relevant for the brain.

Prediction and decision making

Knowledge of the environment as represented by a probabilistic generative model can be utilised for prediction by conditioning on the observed state of the world and computing the distributions over variables describing future states of the environment. Alternatively, the agent can condition not on observed, but possible states and actions to deduce their consequences. The evaluation of such what-if scenarios supports planning and decision making. This computational problem is widely studied in the disciplines of economics, game theory, control theory and machine learning. In the typical formalisation called reinforcement learning, the environment is modelled as a Markov decision process (MDP)[30] or its counterpart with a partially observed state (POMDP). An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is a finite space of possible actions, $\mathcal{T}$ is the transition matrix defining the transition probabilities $\mathcal{T}^a_{ss'} = P(s'|s, a)$ and $\mathcal{R}$ is a reward function specifying the average reward for a given state-action pair $\mathcal{R}(s, a) = \mathbb{E}[r|s, a]$. The

[5]However if it later turned out that the less likely interpretation was the correct one, noticing this could be perceived as funny, which is the basis of the computational account of humour of Hurley et al. [29]

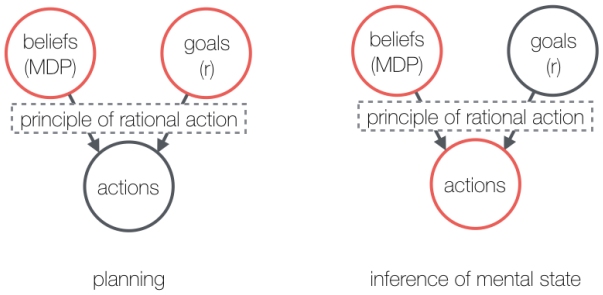


Figure 1.10: Action understanding as inverse planning. Red circles identify the nodes that we condition on, to calculate an estimate for the grey nodes. Figure based on Baker et al., 2019 [31].

agent's decisions are captured by positing a policy that it follows, formalised as the policy function $\pi(s, a) = P(a|s)$. Provided that an agent has learned the correct model of the environment, it can 'solve' the MDP by determining the optimal policy π^* for which the expected return is maximal. This can be thought of as generating an infinite amount of hypothetical what-if situations for both states of the world and actions taken, and evaluating the consequent outcomes. The principle of rational action can then be formalised as the agent following the policy that led to the best outcomes on average, based on its model of the environment (Fig. 1.10).

Similarly to the case of vision and linguistics, having a generative model also helps in inverse inference tasks. We have seen how the forward model combines an agent's beliefs (estimate of MDP and state) with their goals (reward function) to plan actions according to the principle of rationality. What the inverse model enables us to do is to infer an agent's goals given its actions and beliefs about the world, by integrating the likelihood of the observed actions with the prior over goals, which may underlie humans' *intuitive psychology*, also termed a '*theory of mind*' (Fig. 1.10) [31].

1.2.3 Learning a model of the environment

We have seen that semantic memory, in the form of a probabilistic generative model, can effectively support cognitive functions such as perception, prediction and decision making. But how does the brain build such a model of the world based on sensory experience? Sensory experience is difficult to interpret even if the generative model is known due to noise, ambiguity and

resource constraints. Having to also learn the model itself introduces an additional challenge, the problem of induction. Understanding how experiences can be generalised into theories have occupied philosophers of science at least since Hume [32], through Quine, Carnap and Goodman. However, recently these issues have entered the domain of engineering and science as practical difficulties in constructing intelligent artificial agents, where it is commonly known as the problem of generalisation.

A famous illustration of this problem is learning concepts: If we have observed a white swan, and have never observed any black swans, can we justify the statement that all swans are white? What if we have observed a huge amount of white swans and not a single black one [33]? In other words, given a variety of hypotheses about how the world works that are all consistent with observations to date, which one should we choose? The Bayesian approach recognises the similarity of this problem to that of ambiguity in perception and applies the same solution. That is, it views the problem of model induction as the inference of an additional, higher level latent variable that contributes to how the world generates sensory experience. Model inference can be further decomposed into a hierarchy of additional levels such as parameters, structure, structural form and hyperparameters, introducing hypothesis spaces over hypothesis spaces with priors over priors, each level defining a probability distribution over the variables below it [34].

Structure learning in compositional hypothesis spaces

In order to see how hierarchical Bayesian inference can be applied to a concrete problem, consider the example of learning how to make coffee with a new device that one has no prior experience with. Each row in Fig. 1.11, left, summarises observations in a concrete experience of making coffee and tasting it. The variables shown in each column are the results of perception, already incorporating the first step of probabilistic inference in the model. One can describe how each variable such as the type of beans, grind setting or water quality affects the taste using a parametric model and based on the experiences $\mathcal{D}$, the parameters can be tuned to predict the taste for new settings accurately by computing $P(\theta|\mathcal{D})$. However, this parameter estimation process relies on having already identified which variables are potentially relevant, as well as

which of these variables affect which others, for example that the clothes one is wearing are unlikely to have a strong influence on the taste (Fig. 1.11, right). This additional task, called structure learning, can be handled by extending the model with an additional level of hidden variables corresponding to model structure S, so that the generative model becomes $P(\mathcal{D}|\theta, S)P(\theta|S)P(S)$.

Figure 1.11: Causal structure learning. **Left,** Summary of coffee-making experiments, with each row showing the observed values of variables (some variables omitted for brevity). **Right,** Example of a causal model illustrating the relationships between variables and their effect on the quality of the drink. A single estimation sub-problem is highlighted, that of estimating the parameter θ describing the effect of pressure on coffee quality.

Structure learning then relies on evaluating candidate model structures (Fig. 1.12) by computing the posterior over structures

$$P(S|\mathcal{D}) \propto \int P(\mathcal{D}|\theta, S)P(\theta|S)P(S)d\theta.$$

The main challenge in calculating this posterior is that the space of potential model structures becomes vast as the number of variables increases. If the model allows for the addition of latent variables to represent hidden structure in the generative process, then the space of model structures is infinite or open-ended. Furthermore, the space is discrete and the likelihood of each candidate must be evaluated separately, including the marginalisation over parameters. To deal with open-ended model spaces, one approach is to exploit compositionality by defining building blocks and a set of rules on how they can be combined infinitely to produce models of increasing complexity. The building blocks can be for example graphs [35], matrices [36] or programming language primitives [37, 38], and the rules often take the form of generative grammars that define possible replacements of elements of the model with more complex

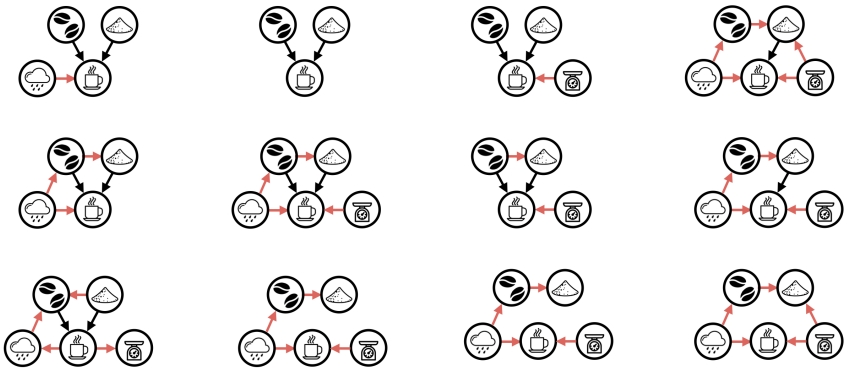


Figure 1.12: Samples of candidate model structures in the coffee making example. Each graph is created by adding or removing edges to and from a simpler candidate structure. The size of the hypothesis space scales combinatorially with the number of variables.

combinations of the primitives.

Kemp et al. [35] uses graphs and a graph growing grammar for expressing a wide range of structural forms such as trees, linear orders, multidimensional spaces, rings, dominance hierarchies or cliques (Fig. 1.13B&D). These forms F are introduced as an additional level in the hierarchy above the inference of the model structure S and they perform inference over both levels

$$P(S, F|\mathcal{D}) \propto \int P(\mathcal{D}|\theta, S)P(\theta|S)P(S|F)P(F)d\theta.$$

Given data from natural domains, their algorithm discovers sensible structures, such as a tree structure for animal features, linear structure for US Supreme court judge voting patterns, ring structure for colour similarities and a cylindrical structure for proximities of cities around the globe. Their algorithm also exhibits developmental shifts in model structure as observations are increased, reminiscent of similar shift in human development in the understanding of categories. Grosse et al. [36] define a context-free grammar on matrix decompositions that can infer latent components and estimate predictive likelihood for nearly 2500 structures and unifies a variety of pre-existing unsupervised learning methods (Fig. 1.13A). Starting from a simple Gaussian matrix, they perform greedy search over the candidate structures produced by the grammar (Fig. 1.13C), using a simpler proxy for the marginal likelihood for scoring them. Their algorithm successfully infers structures in a wide range of natural datasets, including image patches, motion capture data, 20 Questions answers,

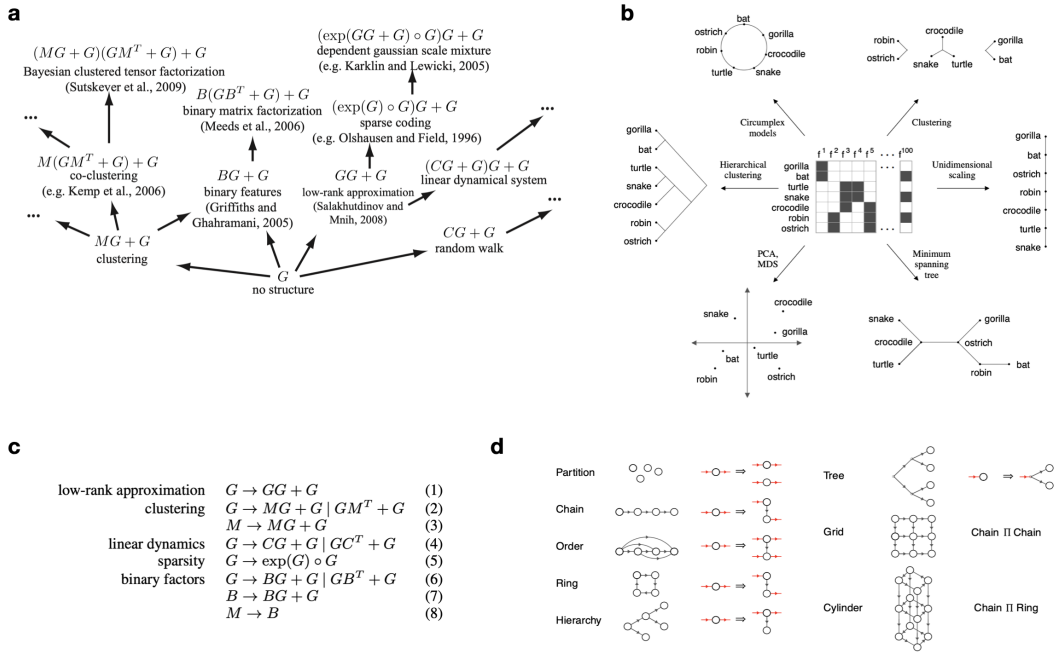


Figure 1.13: Compositional hypothesis spaces. **A,** Exploring the space of model structures through composition of matrices gives rise to previously defined statistical models from the machine learning literature in the framework of Grosse et al. [36]. Arrows indicate models that can be reached by applying a single production rule, but only a subset of possible steps are shown. **B,** Model structures that can be discovered given a matrix of binary features in the method of Kemp et al. [35]. **C,D,** Generative grammars for combining primitives leading to A and B respectively. A, C reprinted from Grosse et al. [36], B, D reprinted from Kemp et al. [35].

and U.S. Senate votes.

Open-ended model spaces can also be constructed by directly exploiting compositionality in the same manner that is used by human programmers to navigate the space of algorithms: that is by constructing programs from the primitive instructions of a general purpose programming language. This approach frames learning as program induction, that is, based on observing the outputs of a program, the task is to infer what could be the code that produced them. Viewing learning as program induction aligns directly with representing knowledge as probabilistic programs that we have introduced in section 1.4, as well as theoretical approaches for defining universal learning agents [39]. It has also been argued by Rule et al. [40] that it might be fruitful to extend the analogy with additional techniques employed by programmers, such as adding and extracting functions, debugging, profiling, refactoring, writing libraries and inventing domain specific languages. As concrete examples,

Piantadosi et al. [37] have used a restricted set of primitives to model the acquisition of the concept of natural numbers by children over the course of their development, based on word statistics from child-directed speech (CHILDES dataset). Ellis et al. [38] built a general purpose program induction algorithm in a wide range of domains such as list processing, text editing, drawing of simple shapes, building block towers, symbolic regression and equation discovery. Their algorithm, called DreamCoder, is capable of constructing domain specific libraries and using its previously defined abstractions as building blocks for future induction tasks. This enables the algorithm to successfully rediscover fundamental building blocks of functional programming, vector algebra, and classical physics, including Newton's and Coulomb's laws.

Common in these approaches that explore a vast space of compositionally defined model structures is that they proceed from simpler models to more complex ones as more data is observed. This progression over models relies critically on a feature of Bayesian model selection called the *automatic Occam's razor*, which automatically trades off between complexity and model fit, protecting against selecting overly complex models. In the following section, we explore how this feature of Bayesian model selection emerges in the computation of the model posterior.

Automatic Occam's razor

Popper observed that there is a fundamental difference between the way we can confirm or falsify hypotheses. He noted that no amount of evidence supporting a hypothesis can definitively prove it, but a single counterexample can disprove it. For example, while even an immense amount of white swans can't prove that all swans must be white, the observation of even one black swan falsifies it. He concluded that we should be suspicious of theories that are not falsifiable [33]. This preference for falsifiability is reflected in Bayesian model selection, known as the automatic Occam's razor effect [41]. The automatic Occam's razor penalises models that can accommodate a wide range of outcomes, relative to models that make more precise predictions.

To understand how Bayesian inference incorporates the automatic Occam's

razor, consider the posterior over models (M) as a dataset ($\mathcal{D}$) is observed:

$$P(M|\mathcal{D}) = \frac{P(\mathcal{D}|M)P(M)}{\sum_M P(\mathcal{D}|M)P(M)}.$$

The posterior over models gives the degree of belief that the agent assigns to each candidate hypothesis. If the agent has no prior preference for any of the considered models, that is to say it uses a uniform prior $P(M)$, then the normalised model likelihood $P(\mathcal{D}|M)$ is equivalent to the model posterior. Therefore, we focus on this quantity, also called the marginal likelihood highlighting that it can be computed by marginalising over the parameter values θ:

$$P(\mathcal{D}|M) = \int P(\mathcal{D}|\theta, M)P(\theta|M)d\theta. \tag{1.2}$$

The quantity that evaluates how well a given model can fit the data is the likelihood $P(\mathcal{D}|\theta, M)$, which is commonly used for selecting the right parameters in statistical inference and machine learning (maximum likelihood estimation). However, it has a critical problem: models differ in their flexibility, and certain models and can achieve high likelihood on almost any data set; in the words of John von Neumann, 'with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.' [42]. This problem is analogous to the problem of falsifiability.

If we rely on Bayesian model selection and compare the marginal likelihoods instead of the standard likelihood, then as we noted in Eq. 1.2, we also have to marginalise over possible parameter settings. One way to interpret the effect of this marginalisation is to rewrite Eq. 1.2 as a Monte Carlo integral, where we take the average likelihood over samples from the parameter priors:

$$P(\mathcal{D}|M) = \lim_{N\to\infty} \frac{1}{N} \sum_{\theta^i \sim P(\theta|M)}^{N} P(\mathcal{D}|\theta, M)$$

This way of expressing the marginal likelihood makes it clear that a model will be preferred according to this measure if it predicts the dataset not with fine-tuned, but with randomly sampled parameters, and consequently, complex models that rely on a precise calibration of their parameters are penalised. A different way to look at the same effect is to observe that complex models, by virtue of being consistent with many different observations, spread their predic-

tive probability over a large volume in the space of possible data sets. Since the predictive distribution is normalised, the absolute value of such widely spread distributions at any given data set has to be low (Fig. 1.14). Consequently, given a simple model with strong, easily falsifiable predictions and a complex model that can explain almost anything with high precision, the marginal likelihood is going to prefer the simpler model even if it is slightly less successful in explaining the data, instantiating the automatic Occam's razor. Note that frequently used model selection criteria in statistics such as AIC and BIC can be derived as approximations of Bayesian model selection under simplifying assumptions.

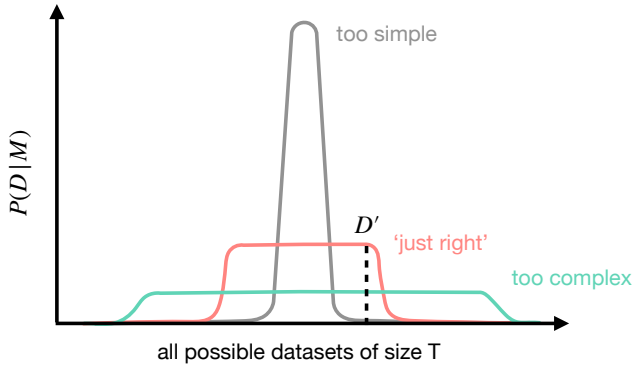


Figure 1.14: The automatic Occam's razor. Model likelihoods over all possible data sets for three models that spread their predictive probability mass in different ways. The blue model can fit many kinds of data, but due to the normalisation constraint this also means that it has to assign lower probability to a given data set than a simple model (green). A given dataset is indicated with the dotted line - we can see that the model likelihood prefers a model that is neither too simple, nor unnecessarily complex. Figure adapted from Rasmussen et al. [43].

Online vs batch learning

A further aspect of learning to consider is that, in contrast to many machine learning settings, the brain does not have access to a lifetime of experiences for training. Instead, it has to incorporate new observations into its model estimate as they arrive. In machine learning, this former setting is called *batch learning*, while the latter is called *online learning*. In the batch learning problem the task is to learn a model from some training set $\mathcal{D} = \{x_1, x_2, ..., x_N\}$ and use this model on a test set. In the Bayesian inference framework this can be viewed

as the inference of the model parameters after observing training set,

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{\sum_\theta P(\mathcal{D}|\theta)P(\theta)},$$

and (optionally) choosing the model that maximises the posterior over the parameters (MAP estimate):

$$\hat{\theta} = \underset{\theta}{argmax} P(\theta|\mathcal{D})$$

In contrast, in an online learning scenario the data points $x_{t_i} \in \mathcal{D}$ arrive individually in succession, and the agent has to refine its predictions after each point. This iterative refinement is ideally regulated by an update rule that depends only on the estimator from the previous time step and the current data point. Formalising this in the Bayesian framework, the posterior after the first data point $\mathcal{D}_{t_1} = \{x_{t_1}\}$ is

$$P(\theta|x_{t1}) \propto \frac{P(x_{t1}|\theta)P(\theta)}{\sum_\theta P(x_{t1}|\theta)P(\theta)}$$

Calculating the joint posterior for $\mathcal{D} = (x_{t1}, x_{t2})$, we may notice that the previous posterior appears in the result:

$$P(\theta|x_{t1}, x_{t2}) = \frac{P(x_{t2}|\theta)\overbrace{P(x_{t1}|\theta)P(\theta)}^{previous\ posterior}}{\sum_\theta P(x_{t2}|\theta)P(x_{t1}|\theta)P(\theta)} = \frac{P(x_{t2}|\theta)P(\theta|x_{t1})}{\sum_\theta P(x_{t2}|\theta)P(\theta|x_{t1})}$$

Intuitively, this means that we can view the previously calculated posterior as the agent's prior in the inference of the subsequent posterior (Fig. 1.15):

$$P_{t+1}(\theta) = P(\theta|\mathcal{D}_t) = P_t(\theta|x_t) = \frac{P(x_t|\theta)P_t(\theta)}{\sum_\theta P(x_t|\theta)P_t(\theta)} \tag{1.3}$$

1.3 Resource constraints

Bayesian inference provides a normative framework for addressing many of the challenges that the brain faces, focusing on a principled treatment of uncertainty and ambiguity. However, it does not address a similarly fundamental challenge: the severe resource constraints under which the brain must operate.

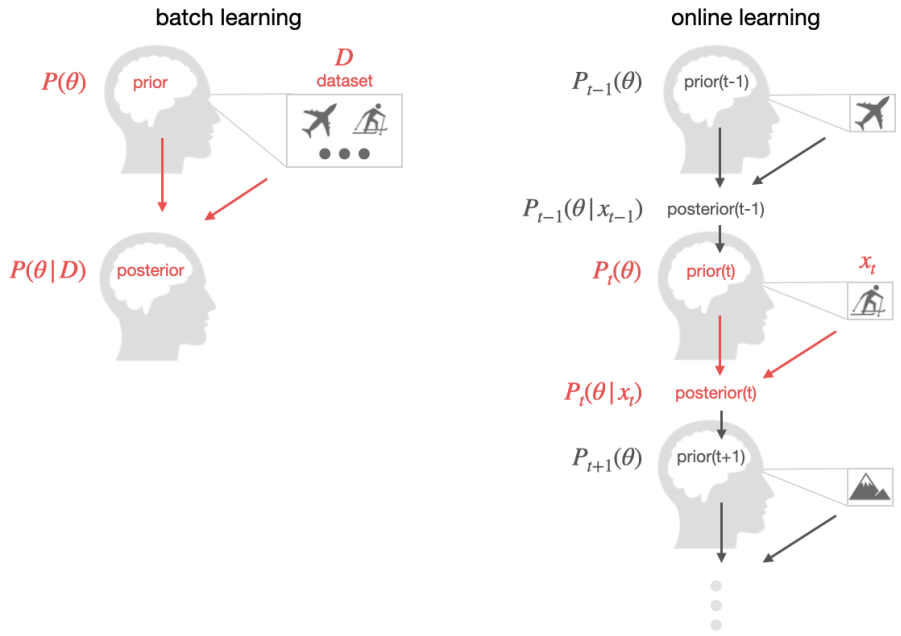


Figure 1.15: Comparison of Bayesian inference for an online learning and a batch learning scenario. **Left:** In batch learning, all observations are available at the same time and the posterior can be computed in a single step. **Right:** In online learning, at each observation, the posterior incorporates all prior knowledge from preceding observations, and this posterior then acts as the prior in the next inference step.

These constraints come in various forms, such as constraints on time available for computation, as well as constraints on the physical size of the brain due in part to the metabolic costs of maintaining and operating it. Bayesian inference, as a computational level framework for dealing with uncertain knowledge, assumes boundless computational power and performing the computations that it prescribes is often not feasible. Consequently, a large variety of approximate methods have been developed and one goal of this section is to offer a very brief overview of a selection of these that I rely on in the rest of the thesis. I note however, that in my opinion, these resource constraints are similarly fundamental as the issue of ambiguity, and therefore would merit an integration into the design of the normative framework for understanding the brain as opposed to the post-hoc remedies that these approximation methods provide.

Perhaps the most influential approach for engaging with resource constraints in a principled way in computational neuroscience and cognitive science is information theory. It formalises information, an abstract measure of representational capacity, and provides a normative framework both for compressing information losslessly as well as discarding it in a principled manner

as representation resources are reduced. Additionally, efficient storage of past observations is a fundamental aspect of the computational problem of memory. Therefore I briefly introduce information theory in this section, with further details in chapter 3 which specifically relies on the theory of *lossy* compression, called rate distortion theory.

1.3.1 Approximate Bayesian inference

In this section, I provide a brief overview of two commonly used approximate Bayesian inference approaches: Monte Carlo and variational inference. Monte Carlo methods use sample-based representations of probability distributions, which become more accurate as the number of samples increases. Variational methods, on the other hand, assume simpler parametric forms for the distributions to be approximated and optimise these parameters to reduce the deviation from the true solution.

Monte Carlo approximations

Instead of representing $p(\theta)$ directly as a function, we can represent it as a suitably large set of i.i.d. samples

$$p(\theta) \approx \frac{1}{N} \sum_i \delta_{\theta^{(i)}} = \{\theta^{(i)} \sim p\}_N \tag{1.4}$$

where $\delta_{\theta^{(i)}}$ is a point mass indicator function that takes the value 1 if $\theta = \theta^{(i)}$ and otherwise 0 and $a \approx b_N$ denotes that $a = \lim_{N\to\infty} b_N$ (asymptotically equal) and $\mathbb{E}[b_N] = a$ (unbiased). Intuitively, the RHS of Eq. 1.4 defines a histogram that tends to $p(\theta)$ as the number of samples is increased. Furthermore, we can use the samples to approximate the expected value of any function over $p(\theta)$ by taking the average of the function over the samples

$$\mathbb{E}_p[f] = \int f(\theta)p(\theta)d\theta \approx \frac{1}{n} \sum_{\{\theta^{(i)}\sim p\}} f(\theta^{(i)}) \tag{1.5}$$

The target of these approximations in Bayesian inference is usually the posterior distribution, however sampling directly from the posterior is often not tractable. One of several approaches to this problem called importance sam-

pling is to take samples from a simpler distribution $q(\theta)$ and correct for the deviation from $p(\theta)$ with suitably chosen $w^{(i)}$:

$$p(\theta) \approx \{w^{(i)}, \theta^{(i)} \sim q\} \tag{1.6}$$

A general way to choose the proposal distribution that is well adapted to an online learning setting is the Sequential Monte Carlo (SMC) method, also called *particle filter*ing (PF). In SMC, we use the prior for obtaining the samples (or particles) and then correct the deviations from the posterior updated with a single additional observation. Interestingly, in this case, the required weights turn out to be the likelihoods of each sample on the newly observed data point, normalised such that the weights sum to unity. Using this weighted sample representation of the updated posterior as the new prior, we continue recursive online updates over the entire dataset as in Eq. 1.3.

$$p_{t+1}(\theta) \approx \{Z^{-1} p_t(x_t|\theta^{(i)}), \theta^{(i)} \sim p_t(\theta)\} \tag{1.7}$$

One of the main issues with SMC/PF is sample diversity: after a large number of updates, the weights of most samples tend to zero. A frequently used mitigation strategy is to periodically resample the representation using the weights as the parameters of a multinomial distribution $\{\theta^{(i)} \sim Multinomial(\{w^{(i)}\})\}$ [44]. This simplest variant of a resampling step makes copies successful particles and removes particles with low weights, and it can be seen as a method for converting a weighted sample representation of a distribution into a standard sample representation format (Fig. 1.16).

SMC/PF was originally developed for estimating the posteriors of state variables such as the latent variable in a hidden Markov model. In such models, the latent dynamics increases diversity by separating the resampled copies of particles. However, if we are computing a posterior over a static parameter, the original prior samples never change their locations, only their weights. This can be mitigated by introducing artificial dynamics, or combining with different sampling methods such as Markov Chain Monte Carlo (MCMC), although such solutions are technically complex and their convergence properties are not extensively studied [45].

The second important drawback to standard SMC/PF methods is that

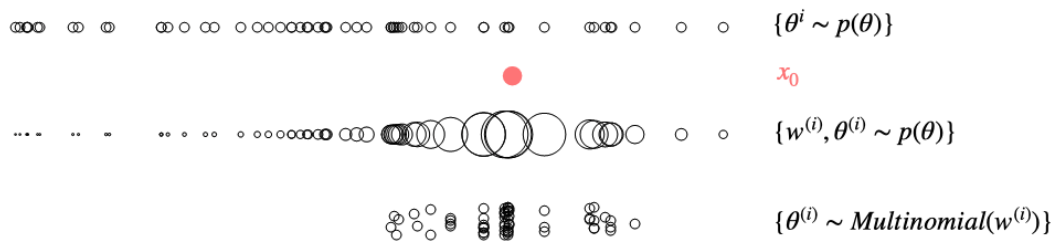


Figure 1.16: Schematic illustration of sequential importance sampling. Each circle represents a weighted sample, where the center of the circle represents the sample value and the radius of the circle is proportional to the sample's weight. First, the learner generates uniformly weighted samples from the prior (first row). Then, it observes the point x_0 (second row) and weights the prior samples using the likelihood, resulting in a weighted sample representation of the posterior (third row). The posterior can then be converted into an unweighted representation through resampling (fourth row) and this can serve as the prior for incorporating a subsequent observation.

they do not scale well to high dimensions. An efficient mitigation technique is available when a subset of these dimensions can be treated analytically then each particle can carry an associated analytical distribution and the effective dimension of the particle filter can be reduced. This method is called the Rao-Blackwellisation of the sampler [46].

Variational inference

Variational inference (VI) algorithms constitute the main alternative to Monte Carlo methods for performing approximate Bayesian inference. Variational methods use a parametric distribution family $q_\phi(\theta)$ and find the element of the parametric family that lies closest to the true solution in some measure (Fig. 1.17). The most commonly used measure is the KL divergence, which we introduce in section 1.3.2, resulting in the objective

$$\text{argmin}_\phi \, \text{KL}(q_\phi(\theta|x) \,||\, p(\theta|x)). \tag{1.8}$$

Since this form of the objective contains the intractable posterior, it is not directly applicable. However, using the definition for the KL divergence and reordering the terms, we can get the following expression:

$$\mathbb{E}_{q_\phi(\theta|x)}[\log p(x,\theta) - \log q_\phi(\theta|x)] = -\text{KL}(q_\phi(\theta|x)||p(\theta|x)) + \log p(x),$$

The LHS of the equation is called the evidence lower bound (ELBO) or the

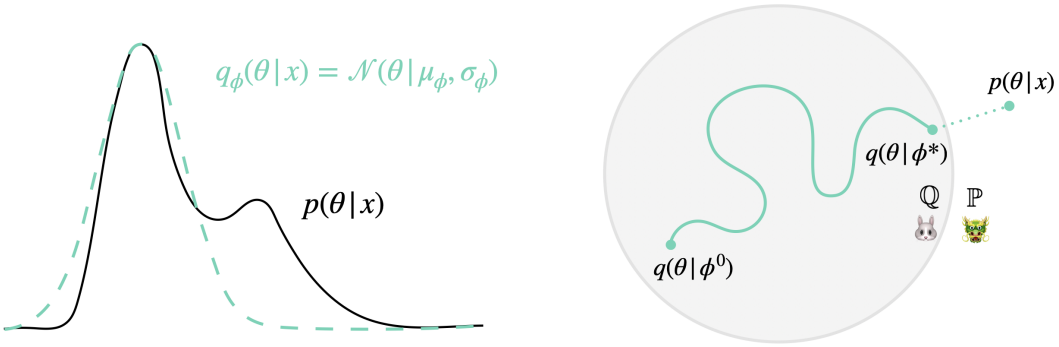


Figure 1.17: Schematic illustration of variational inference. **Left,** a more complex distribution (black line) is approximated using a simpler, parametric distribution (green dashed line). **Right,** the posterior $p(z|x)$ defines a point in the space of all probability distributions $\mathbb{P}$. The approximating parametric family of distributions $\mathbb{Q}$ define a small subspace within $\mathbb{P}$ (emoji notation from Huszár, 2017 [47]). The goal of VI is to find the point $q(z|\phi^*)$ in $\mathbb{Q}$ that lies closest to the true posterior by optimising ϕ.

negative variational free energy. Since the KL divergence must be positive, the ELBO is a lower bound on the log marginal likelihood $\log p(x)$. Also visible from this equation, maximising the ELBO is equivalent to minimising the KL divergence on the RHS, allowing us to optimise $q(\theta|x)$ without having access to the true posterior.

Amortised inference

The idea of amortising Bayesian inference has recently been proposed as an additional idea for mitigating resource constraints in human cognition [48]. In general, amortisation trades off memory resources to save on computational costs by storing the results of previous computations in order to make future ones more efficient [49]. For example, in Monte Carlo methods, the recursive sequential updates we have introduced in Eq. 1.7 reuse the samples from the previous step and therefore can be seen as a form of amortisation.

Another successful variant, deep amortised VI, combines the ideas of amortisation and VI with advances in deep learning. In this approach, we perform variational inference to compute the posterior over state variables on data from a training set. At the same time, we train a deep neural network on these data-posterior pairs, treating the mapping from data point to posterior $f : x_t \mapsto q(z|x_t)$ as a supervised learning problem. In this view, the role of the neural network is to provide a parametric function family f_ϕ with good gener-

alisation properties. After training, the output of the network offers an easy to compute, amortised approximation of the posterior for novel data points. This approach is adopted in one of the most widely used generative models called variational autoencoders and its variants, which we build heavily on in our formalisation of the interactions of the episodic an semantic memory systems. For a brief introduction to variational autoencoders see section 3.1.3.

1.3.2 Information theory

From the perspective of an agent, information from the environment means that the environment produces one of a certain number of possible alternative observations that the agent may expect. The quantity of information that an agent receives depends both on the number of alternatives and to what degree the agent expects the occurrence of the observation. If an observation x has a probability of $P_{agent}(x) = 1$, it contains no new information for the agent as it is certain to occur. Conversely, the lower the probability of the message, the greater the information it holds, i.e. the more surprising it is. A second, mathematical desideratum for information is that it is additive for independent sources. For example for two independent sources X, Y the information content should be the sum of their individual information contents: $I(x \cup y) = I(x) + I(y)$. The simplest way to satisfy these two intuitions is to define information content of an outcome $I(x)$ as [50]:

$$I(x) = \log \frac{1}{P(x)}$$

measured (depending on the base of the logarithm) in bits or nats. The uncertainty inherent in a particular information source is quantified by the *entropy*, which is the average information content or average *surprise* of the source.

$$H(x) = \mathbb{E}_P[I(x)] = -\int_{\mathbb{R}} P(x)\, \log P(x)\, dx,$$

measured in the same units as information content. Entropy is maximal if all outcomes are equally probable. If an agent has a model of the world Q that differs from the true generative process P and therefore leads to incorrect expectations, the average surprise that the agent experiences is quantified by

the *cross entropy*:

$$H_{\times}(P,Q) = \mathbb{E}_P[I_Q(x)] = -\int_{\mathbb{R}} P(x)log(Q(x))dx.$$

In the case of $P = Q$, this is equal to the entropy of P. We can compute how much of the uncertainty is a consequence of using the wrong model by subtracting the inherent uncertainty of P, leading to the concept of *relative entropy* or Kullback-Leibler divergence:

$$\mathrm{KL}(P,Q) = H_{\times}(P,Q) - H(P) = \mathbb{E}_P\left[\log\frac{P(x)}{Q(x)}\right], \tag{1.9}$$

It can be easily seen that $\mathrm{KL}(P,Q) = 0 \Leftrightarrow P = Q$ and Gibbs' inequality guarantees that $KL(P,Q) \geq 0$, making it a semi-quasimetric. Hence, it can be thought of a kind of distance in the space of probability distributions, keeping in mind that it is not a true metric as it is neither symmetric ($KL(P,Q) \neq KL(Q,P)$) nor sub-additive. The KL divergence is a central quantity in machine learning and statistics. For example, it can be shown that the maximum likelihood method in statistics minimises the KL divergence between the observed *empirical distribution* and the predictive distribution,

$$\arg\min_{\theta}\ \mathrm{KL}\left[\, P_{data}(x)\,,\, P_{\theta}(x)\,\right] \longleftrightarrow\ \arg\max_{\theta}\log P(x|\theta)$$

as well as provides the optimisation objective for variational inference, see section 1.3.1 and section 3.1.3.

A further important quantity is mutual information, $I(X,Y)$, which measures how much an agent learns about random variable X from observing random variable Y

$$I(X,Y) = H(X) - H(X|Y).$$

Mutual information (MI) can be seen as a measure of the dependence between the two variables, and it can also be expressed as the KL divergence between the joint distribution of X,Y and the product of the marginals,

$$I(X,Y) = \mathrm{KL}[P(X,Y)\,||\,P(X)P(Y)]$$

which is zero if and only if X and Y are independent. MI is a central quantity in the information bottleneck method [51] that we rely on for formalising how

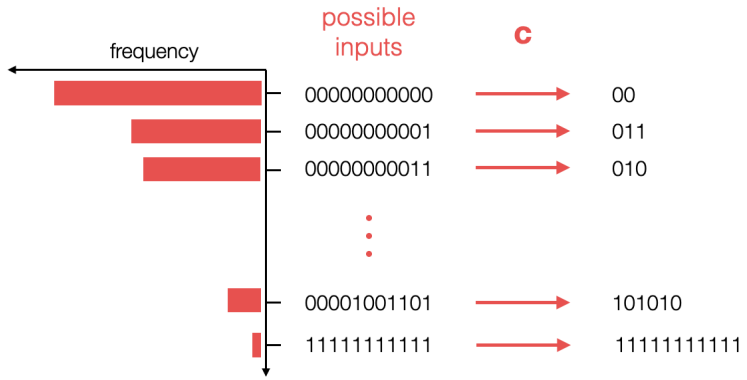


Figure 1.18: Schematic illustration of lossless compression and the Kraft-McMillan theorem. In the middle column, we see possible binary inputs from the source, with the relative frequency of each input shown as a histogram on the left. The coding function c maps each input to a binary code, shown in the right column. The underlying idea in lossless compression is that the coding function maps short code words to more frequent inputs.

semantic memory supports the efficient storage of episodic memories in chapter 3.

Information theory and memory

Mutual information can also be used to capture the notion of *predictive information*, that is the information in past observations $\mathbf{x}_P = \mathbf{x}_{0:t}$ that is relevant with regards to predicting future sensory input $\mathbf{x}_F = \mathbf{x}_{t+1:T}$ [52]. It can be shown that if the environment is a generative process $P(x|\phi)$ that the agent models with $Q(x|\theta)$, then the information that the agent has regarding future observations is exactly the information that its model contains regarding the true generative process [9], that is,

$$\lim_{T\to\infty} I_{pred} = \lim_{T\to\infty} I(\mathbf{x}_P, \mathbf{x}_F) = I(\theta, \phi)$$

This lends formal support to our intuitive argument in the introduction that if the function of memory is to support predictions, then the brain should construct an approximation of the generative process that produces its observations.

A second connection between information theory and memory can be drawn if we consider that memory has to store past experiences. Efficient storage entails an encoding of observations into a shorter form, and according to Shannon's source coding theorem the fundamental limit to such compressed

description of experiences is the entropy of the information source that is to be encoded. Specifically, the theorem states that if we want to encode a random variable X taking values from $\mathcal{A}_X$ with entropy $H(x)$ using an alphabet $\mathcal{A}$, there exists an optimal, uniquely decodable code for which the expected length of code words is bounded by the entropy:

$$\frac{H(x)}{\log|\mathcal{A}|} \leq \mathbb{E}_P[|c(x)|] < \frac{H(x)}{\log|\mathcal{A}|} + 1,$$

where an encoding is a mapping $c : \mathcal{A}_x^* \to \mathcal{A}^*$, with $*$ being the Kleene-star $\mathcal{A}^* = \bigcup_{i\in\mathbb{N}} \mathcal{A}^i$. The idea underlying a compressed description of observations is to attach shorter codes to likely observations, at the cost of lengthening the description of unlikely ones. The mapping between code lengths and the probability distribution of the information source is established by the Kraft-McMillan theorem, stating that analogously to the normalisation constraint on probabilities, the sum of code word lengths cannot be arbitrarily small and for uniquely decodable codes, their sum has to satisfy

$$\sum_{i=1}^{n} \left(\frac{1}{|\mathcal{A}|}\right)^{\ell_i} \leq 1,$$

where $l_i = |c(x_i)|$ is the length of the code for x_i in bits. Intuitively, there are exponentially more long words than short words, and so the cost of using a code word of length l_i scales as $|\mathcal{A}|^{-l_i}$. Therefore if we want to optimally allocate the code word budget, we should choose them such that $P(x_i) = |\mathcal{A}|^{-l_i}$ (Fig. 1.18). In this sense, any lossless compression algorithm can be seen as representing an implicit probability distribution, and conversely, the brain has to have at least an implicit probability distribution over possible experiences if it is to use its memory capacity efficiently. An interesting way to think of this is that the probabilistic model in semantic memory generates a language (coding scheme) to which raw sensory data is translated, enabling concise descriptions of past, present and possible future (or even imaginary) states of the world. Note that the relevant compression regime for the brain is more likely to be the lossy one, which requires an extension of the framework to handle distortions. The application of this extended framework, called rate distortion theory, to human memory, is the main subject of chapter 3.

1.4 Kinds of memories

While there is no universally agreed upon definition for what constitutes a memory system and how to differentiate them, it is widely recognised that there are multiple memory systems. This understanding is largely derived from observations of patients with various memory impairments, such as amnesia, who exhibit selective losses of certain memory-related cognitive abilities without affecting others. In this section, I will present a commonly accepted classification of memory systems, based on the framework outlined in [53], which serves as a useful reference point.

Sensory memory retains sensory information for a short amount of time after the stimulus has ceased. This can be seen in the classic example of a sparkler in a dark room appearing to leave a trail, which quickly disappears, demonstrating the limited time frame in which this type of memory operates. This ability to retain sensations is believed to exist across all sensory modalities, including hearing (echoic memory) and touch (haptic memory). The capacity to retain small amounts of information for just a few seconds is referred to as *short-term memory*. This is an active form of memory, which can be sustained by conscious effort, such as repeating a phone number repeatedly to remember it. Short-term memory is often seen as part of a larger system, referred to as working memory, which acts like a mental scratchpad for thinking and reasoning. Long-term memory, on the other hand, refers to the cognitive processes that allow for the storage of information over longer periods of time, including hours, days, years, and even decades. It is divided into two types: *declarative* (or *explicit*) and *non-declarative* (or *implicit*) memory.

Implicit memory refers to knowledge that is not easily accessible through conscious thought and is tied to specific actions, such as the skill of riding a bike or a conditioned response of blinking in response to a tone after it has been consistently followed by an air puff. Performance in these kinds of tasks can be improved with practice even in amnesic patients. In contrast, declarative memory encompasses information that can be consciously retrieved, such as general knowledge about the world or personal experiences.

Explicit memory is further divided into two categories: semantic and episodic memory. Semantic memory, as we have discussed, refers to general

knowledge of the world such as concepts or the meanings of words. It also includes integrated chunks of knowledge termed *schema*s, exemplified by a knowledge of how a typical restaurant visit plays out (schemas that describe typical events are called *scripts*), how to book a theater seat or what a typical home consists of (*frames*, which store information about objects and their properties). Episodic memory, on the other hand, relates to our capacity to recall unique and personal experiences. One notable aspect of episodic memory is what Tulving refers to as 'mental time travel', the subjective feeling of reliving the remembered experience. The distinction between episodic and semantic memory is often made based on the subjective feeling of 'remembering' versus 'knowing'.

1.5 Outline

In this thesis, I focus on a normative account of long-term memory, and specifically investigating the interactions between the semantic and episodic memory systems. While a generative model of the environment has been shown to be crucial for adapting to environmental demands, the rationale behind an episodic memory system is less clear. In Chapter 2, I aim to demonstrate that an episodic memory system is necessary for a learning agent and how it complements the semantic memory system. This chapter is based on Nagy & Orban (2016) [54]. In Chapter 3, I examine the opposite perspective, investigating how semantic memory can aid the functioning of episodic memory. Specifically, my goal is to evaluate whether the generative model stored in semantic memory can serve as the basis for the effective compression of sensory experience. However, such lossy compression inevitably introduces distortions in reconstruction, and I also compare these distortions with the artefacts that result from the application of traditional compression algorithms in computer systems and systematic memory distortions identified in prior human experiments. This chapter is based on Nagy et al. (2020) [55], also incorporating results from Nagy et al (2019) [56] and Fráter & Nagy et al. (2022) [57].

In Chapter 4, my goal is to extend the prior computational and algorithmic level analyses to the question of how these algorithms can be implemented in the brain and artificial neural networks. In the first of the final set of studies, we

investigate whether the semantic compression idea can be implemented using hierarchical generative models and contrast qualitative features of neural activities in the model with neural recordings from the visual cortex of macaques. This section is based on results published in Bányai & Nagy et al. (2019) [58]. In the final study, I revisit the sensitivity of both human and machine learning to the order in which observations are presented, which were analysed in the first study. The goal is to examine whether task representations can be maintained separately in the same neural network, such that the artificial neural network shows similar robustness to blocked training as humans. This final section is based on work done during an internship that led to the publication Flesch & Nagy (2023) [59].

Chapter 2

What use is an episodic memory?

In a complex, structured environment that is capable of providing a practically infinite variety of possible experiences, storing them in all their detail would take a prohibitive amount of memory resources and would be useless in responding to novel situations. It is more beneficial for an learning agent to extract the structure of the world into a concise model, which enables both compression and generalisation, and store this model instead of the observations. But then what is the benefit of devoting precious mental resources to encoding inconsequential contingencies by storing rich snapshots of actual experience, that is, what use is episodic memory?

Here, we argue that online learning in open-ended hypothesis spaces (Section 1.2.3) under realistic resource constraints — similar to what the human brain faces — presents a computational challenge that makes such a memory system necessary. In an online learning scenario, observations arrive sequentially and predictions have to be continuously updated. Iterative updates of a particular model's parameters do not require storing the data, since it is sufficient to retain only the information relevant to the specification of the parameters. However, if the structural form of the model is a priori unknown [35], then only a subset of candidate models can be tracked at any given time, since the memory cost of retaining even such compressed statistics becomes prohibitive for an infinite set of models. The inevitable information loss resulting from this restriction presents the brain with a delicate problem: relevance judgements, that is, decisions about what to forget and what to remember can only be based on the currently tracked models. At the same time, the initial

guess for which models these should be is likely to be wrong because the initial data will only warrant an overly simple model and because it might be misleading about the correct structure and form. Introducing such a bias in the interpretation of new experiences towards the wrong models means that statistical power required for model updating cannot accumulate, since the evidence for alternative models and the information needed for fitting those models will often be deemed irrelevant and discarded, preventing the discovery of the correct representation.

We propose that an episodic memory can alleviate the fundamental problem of online learning described above, by retaining a selected subset of samples. This mini-batch allows evidence for a novel model to accumulate by retaining the contingent details of observations irrespective of how relevant they appear under the current model. We also argue that to take full advantage of episodic memory, its contents should be chosen selectively, so that the combination of episodic and semantic memories provide an efficient representation of the observations.

We are aware of two prior attempts to provide a normative explanation for an episodic memory based on computational principles. The complementary learning systems account of McClelland et al. (1995) [60] suggests that a fast, hippocampal learning system is required in order to avoid interference with knowledge stored in a neocortical system where learning occurs via slow changes of synaptic connectivity in a network of neurons. Catastrophic interference can be seen as a special case of the detrimental consequences of an inability to maintain a lossless representation of observations during learning, but in contrast to our treatment, the complementary learning systems approach lacks a normative framework and only concerns parameter estimation within a single model. Lengyel & Dayan (2009) argue that using the data samples directly for control is advantageous at the early stages of learning in a new environment. A different but related question about how the combination of semantic and episodic memories can be used to optimise reconstruction is explored by Hemmer et al. [11]. Mahr & Csibra (2017) [62] ask a similar question and locate the function of episodic memory in its role in human communication, however they concentrate specifically on its metarepresentational structure as distinguished from its representational contents. In contrast, our

work focuses on the latter, specifically on why a large set of irrelevant features are retained within the memory trace, which Mahr and Csibra refer to as event memory.

While this study is intended primarily as a normative argument for the existence of a cognitive system, the problem explored here is intimately related to the efforts in machine learning to handle the problem of online Bayesian model selection in arbitrarily complex model spaces. There are numerous proposals for methods that deal with online model selection or model selection in infinite model-spaces [36] separately. Recently, there have been attempts to tackle both challenges at once in a similar setting, but these are concerned with a restricted hypothesis space over possible model forms, such as mixture models [63, 64, 65]. Methods that are specific to a given model form have the potential to be vastly more efficient within their domain, but we are striving to find the principles for a general purpose computational architecture that is flexible enough to accommodate uncertainty in the structural form of the model [35]. To the best of our knowledge, such a scenario has not yet been explored.

2.1 Learning paradigm

In this study we aim to study how the computational problem of learning shapes the architecture and dynamics of long-term memory. We assume that the main goal of human learning is the acquisition of a suitable representation of the world and propose that this learning process is characterised by the following fundamental properties: i) it is incremental; ii) it requires an open-ended hypothesis space which incorporates not only an arbitrary amount of complexity but also enables the discovery of the appropriate model form; and iii) it is subject to computational constraints, most notably a limited amount of memory.

Our main argument is agnostic to the choice of learning method, but we are adopting the Bayesian inference framework. This framework provides us with a consistent, general and arguably elegant solution for dealing with uncertainty during learning and is central to many state-of-the-art advances in machine learning [66] while simultaneously being able to capture a large body of knowl-

edge concerning the acquisition of abstract knowledge in humans [34, 67]. In this framework the problem of learning can be formalised as the continual refinement and updating of a probabilistic generative model, where information about unobservable or currently not observed variables, parameters and candidate world structures can all be expressed as probability distributions over latent variables (Section 1.2.3).

In our treatment the memory constraints are formalised such that after the model has been updated, by default, the observation is discarded and only the sufficient statistics for the best performing model is kept. The two main challenges introduced by these constraints are that the learner needs to both: i) assess the plausibility and ii) approximate the right parameter settings of alternative models based solely on the sufficient statistics of the tracked model, without having access to the data.

We set out with an example learning problem that can demonstrate both the challenges and the power of the proposed approach: a mixture of Gaussians model (MoG) has the benefit of showing non-trivial model-learning dynamics while also providing an opportunity for analytical treatment. Mixture models are also frequently used as cognitive models of human category learning [68]. We use a version where model selection corresponds to determining the correct number of mixture components based solely on the data; parameter learning consists of finding the means for the components; while mixture weights and variance of mixture components are assumed to be fixed and known. Although a more flexible model would provide richer dynamics, the main challenges stated earlier can be clearly demonstrated on this simplified model.

The rest of the section is structured as follows: first, we show how incremental Bayesian inference works in a setting without resource constraints; next, we introduce a learning agent that only has access to a semantic memory and demonstrate that it has a propensity to discard the information that would enable model change; finally, we show that the introduction of an episodic memory substantially mitigates this problem.

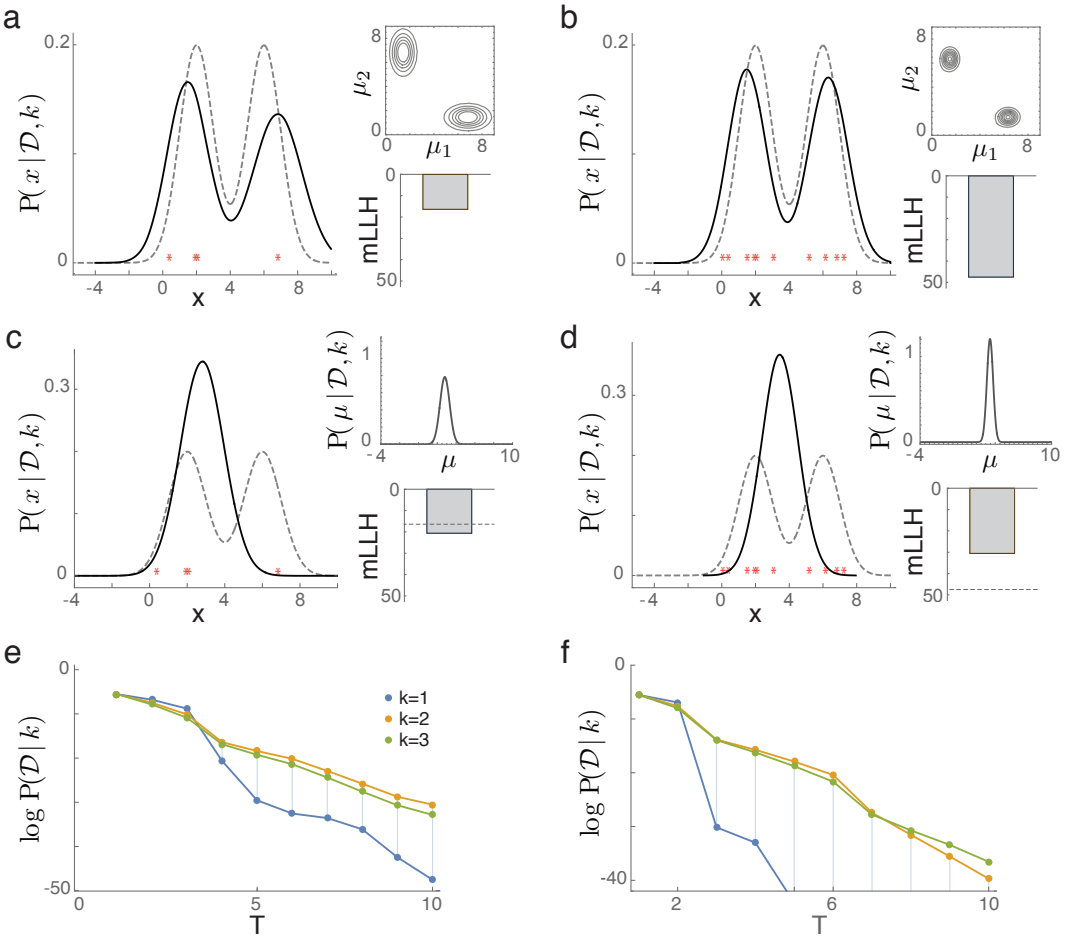


Figure 2.1: Illustration of model learning on a MoG model. **A**, The goal of learning is the estimation of the probability distribution of the data (*left panel, dashed grey line*) from a limited sample (*asterisks*, $n = 4$). Inference in a given model yields a posterior probability distribution over model parameters (*upper right panel*). The model assumes two mixture components ($k = 2$). Based on the posterior, the predictive posterior distribution (*solid black line*) provides our estimate on how data points are distributed. Marginal likelihood assesses the statistical power of the model (*lower right panel*). **B**, Same as **A** but using a larger data set ($n = 10$). Tighter posterior results in a tighter and more accurate predictive probability distribution and higher average marginal likelihood. **C**, **D**, Same as **A** and **B** but for a $k = 1$ model. **E**, Evolution of mLLH as more data is accumulated from a $k = 2$ model. Colours show models with different number of mixture components. Equality of mLLH at $T = 1$ is a consequence of learning limited to the means. **F**, Same as **E** but for a data set from a $k = 3$ mixture.

2.2 Learning in an unconstrained setting

Bayesian inference provides a consistent framework for learning the form, the structure and the parameters of the model estimating the probability distribution of data. Learning entails the estimation of the posterior probability

of parameters (θ) in a given model and/or that of the model (m) itself:

$$\mathrm{P}(\theta \,|\, \mathcal{D}, m) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, \theta, m)\,\mathrm{P}(\theta, m) \tag{2.1}$$

$$\mathrm{P}(m \,|\, \mathcal{D}) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, m)\,\mathrm{P}(m) \tag{2.2}$$

Posterior probabilities for alternative model structures, and/or forms need to be assessed individually and the marginal likelihood (mLLH),

$$P\left(\mathcal{D} \,|\, m\right) = \int d\theta\, P\left(\mathcal{D} \,|\, \theta, m\right) \mathrm{P}\ \left(\theta \,|\, m\right), \tag{2.3}$$

plays a critical role in comparing these models: even with a uniform prior probability distribution over alternative models, the mLLH function implements the automatic Occam's razor principle, which ensures that the simplest model that can account for the observed variance in the data has the highest posterior probability. When the model prior is flat, the evaluation of mLLH is sufficient to compare the models (for more details, see Section 1.2.3).

In the analytic treatment of MoG, the posterior over the means μ is a MoG again, in which the number of mixture components grows exponentially with the number of observations T (for more details see Appendix A.1.1). Whether learning is performed on the whole batch of data at once or is done in an online manner, Bayesian inference yields the posterior distribution of parameters for any particular model structure at any particular time (Fig. 2.1a-d). This posterior distribution can be used to make predictions on upcoming data and learning helps to disentangle the predictions of different models. While early in the training a complex model that reflects the actual statistics of the data adequately might be discounted because of lack of sufficient evidence, after extended experience the marginal likelihood of the simpler model will be overcome by the model of right complexity (Fig. 2.1a-d). Switching time in model selection is determined by the actual data samples and is defined by the evolution of the mLLH (Fig. 2.1e,f). The Automatic Occam's razor that is implemented by the mLLH function ensures that no overfitting happens: the learner discovers more complex structures if data statistics justifies such a model but keeps the model as simple as possible.

2.3 Semantic-only learner under constraints

While Eq. 2.1 provides a general recipe for adjusting the model parameters to data, learning can be formulated in two markedly different ways. i), In order to obtain a posterior at a particular time T, the whole data set $\mathcal{D}^T$ is evaluated according to Eq. 2.1. ii), Online learning relies on a parameter posterior obtained at an earlier time point $T-1$ to provide a prior for the evaluation of novel data:

$$\mathrm{P}\big(\theta \,|\, \mathcal{D}^T, m\big) \propto \mathrm{P}\big(x^T \,|\, \theta, m\big)\, \mathrm{P}\big(\theta \,|\, \mathcal{D}^{T-1}, m\big) \tag{2.4}$$

While online learning has the same power as batch learning, it has the benefit that it is explicitly formulated such that the effect of the earlier data points is summarised in the posterior calculated for $\mathcal{D}^{T-1}$. As a consequence, online learning liberates us from the need to retain the whole data set: once the posterior has been updated the data can be discarded. As long as both parameters and models are updated, this procedure provides a consistent method to update and compare alternative hypotheses on how the model was generated without needing to keep a growing data set in memory. In contrast, if we track only a limited number of models (one model being an extreme but valid approach), discarding data prevents the consistent assessment of alternative models.

The unavailability of the original data leads to an uncertainty as to the possible past data sets that could lead to the same available statistics. An ideal learner represents this uncertainty by means of a probability distribution over possible past data sets. The learner needs a method for constructing such a distribution based solely on the posterior of the current model, since this contains all the information that it has retained. Given such a distribution, a method is required to compare alternative models (i.e. estimate the mLLHs, Eq. 2.3) and to assess what the parameters of the alternative models would have been had those been tracked from the beginning (i.e. estimate parameter posteriors of novel models Eq. 2.1). We propose that a natural approximation of the current model's estimate of the distribution of possible past data sets

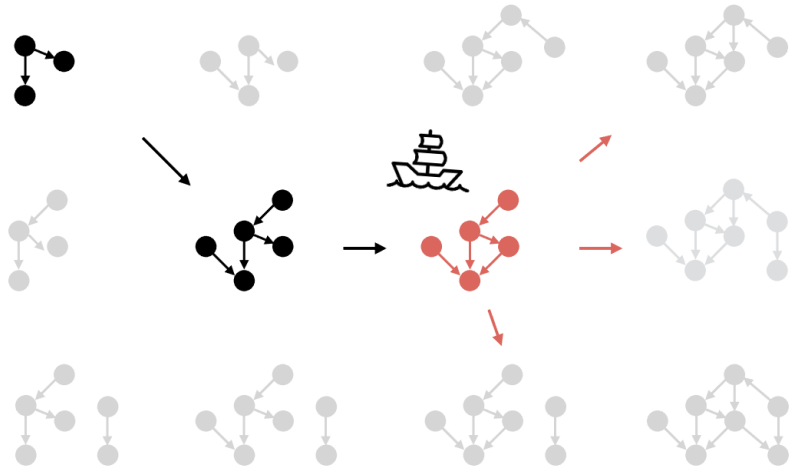


Figure 2.2: Neurath's ship in structure space. The semantic learner can be viewed as maintaining a MC representation using a single sample or particle in hypothesis space that is updated with the arrival of each observation. Such an approximation has also been proposed by Bramley et al. [70] for characterising human learning of causal structure. They repurpose an analogy that was originally drawn between rebuilding a ship while at sea and theory building in science. The original analogy according to Neurath: '*We [theorists] are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.*'

can be obtained by the assessment of the posterior predictive distribution,

$$\mathrm{P}(x \,|\, \mathcal{D}, m) = \int d\theta \, \mathrm{P}(x \,|\, \theta, m) \, \mathrm{P}(\theta \,|\, \mathcal{D}, m) \,,$$

of the tracked model. This choice is conceptually related to using 'pseudopatterns' to transfer knowledge between different models [69]. It has the benefit that while the parameter posteriors of different models in general span very different spaces and are thus not comparable, all models give predictions over the same data space (Fig. 2.1a-d). Another benefit is that the predictive distribution is presumably available for the learner in any case, since it is a fundamental component of numerous other cognitive computations as well.

2.3.1 Inferring the posterior of a novel model

In a given model, the posterior distribution of parameters summarises the model's knowledge about the statistics of the data. Since the predictive distri-

bution of the tracked model carries information about the uncertainty of the parameters this can be used to approximate the posterior of the parameters in a novel model by minimising the dissimilarity of the predictive posterior distributions. Minimising the KL divergence solves exactly this problem:

$$\mathrm{P}(\theta \,|\, \mathcal{D}, m') \approx \underset{\mathrm{P}(\theta|\mathcal{D},m')}{\operatorname{argmin}} \mathrm{KL}\left[\mathrm{P}(x \,|\, \mathcal{D}, m) \;||\, \mathrm{P}(x \,|\, \mathcal{D}, m')\right]. \tag{2.5}$$

Calculating the KL divergence analytically is in most cases unfeasible, therefore two approximations have been made. First, inspired by Snelson, 2005[71] we were looking for a compact representation of the predictive posterior, but instead of achieving this by simply taking a likely set of parameter settings, we've assumed that the posterior comes from a simple parametric distribution family:

$$\mathrm{P}(\theta \,|\, \mathcal{D}, m') \to \mathrm{P}(\theta|\eta, m'), \tag{2.6}$$

where η provides a parametrisation of the approximate posterior. As a result, the former functional optimisation problem in (Eq. 2.5) reduces to

$$\hat{\eta} = \underset{\eta}{\operatorname{argmin}} \mathrm{KL}\left[\mathrm{P}(x|\mathcal{D}, m) \;||\, \mathrm{P}(x|\eta, m')\right], \tag{2.7}$$

where $\mathrm{P}(x|\eta, m') = \int d\theta\, \mathrm{P}(x|\theta, m')\, \mathrm{P}(\theta|\eta, m')$ is the approximate predictive posterior distribution. Eq. 2.7 is equivalent to minimising the cross entropy, which can be approximated using a Monte Carlo integral. After sampling $\hat{x}_i \sim \mathrm{P}(x \,|\, \mathcal{D}, m)$ we have to choose the η for which the expected value of $\log\left(\mathrm{P}(x|\eta, m')\right)$ is maximal, concluding to a maximum likelihood estimation over the generated 'fake data'

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_i \log\left(\mathrm{P}(\hat{x}_i|\eta, m')\right) \tag{2.8}$$

The resulting $\mathrm{P}(\theta|\hat{\eta}, m')$ is our estimate of the parameter posterior on the original data. The parameter posterior in our implementation of MoG is a MoG again, and we have found that a convenient and effective parametrisation of the posterior uses a single mixture component. This approximate posterior effectively reproduces both the true posterior and the true predictive posterior distribution of the model (Fig. 2.3).

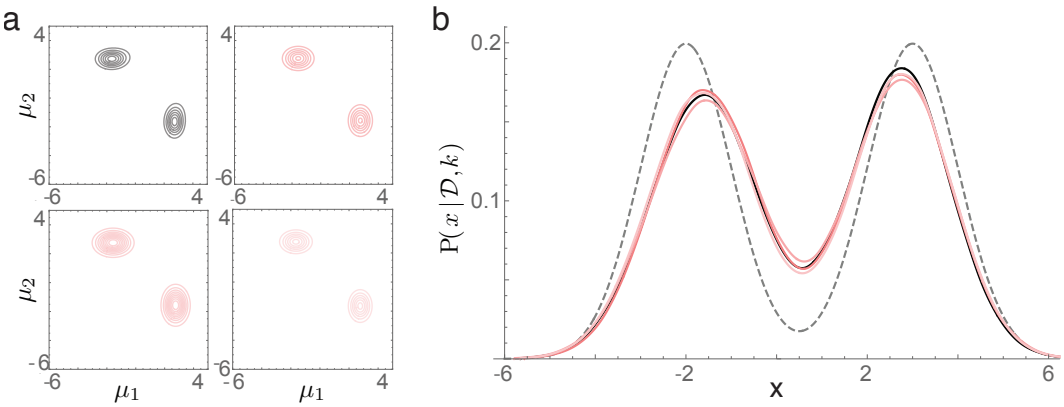


Figure 2.3: Reconstruction of the parameter posterior from the predictive posterior distribution. **A,** Posterior distribution of the component means in a $k = 2$ model after observing $n = 10$ data points. *Black contour plot*: posterior obtained by analytical calculation; *coloured contour plots*: posterior reconstructions. **B,** Comparison of the true predictive posterior distribution (*black line*) and its approximations. Colours are matched across panels, *dashed line*: data distribution.

2.3.2 Model comparison in constrained learners

Model comparison requires the assessment of the mLLH function for alternative models (Eq. 2.3). However, even if we have access to the marginal likelihood of the tracked model, discarding the original data points renders the construction of the mLLH for the novel model impossible. Again, the posterior of the tracked model summarises our knowledge of the data and and therefore we rely on the predictions that can be drawn from the model posterior in order to assess the possible data sets. This can be achieved by calculating the expected value of the marginal likelihood over the predictive posterior:

$$\mathbb{E}_{\mathcal{D}^* \sim \mathrm{P}(\mathcal{D}^* \mid \mathcal{D}, m)}[\mathrm{P}(\mathcal{D}^* \mid m')], \tag{2.9}$$

where $\mathcal{D}^*$ denotes fake data sets obtained from the predictive posterior distribution. This expected value can be evaluated by Monte Carlo sampling. Upon the arrival of a novel data point x^T, fake data sets are sampled from the predictive distribution. The novel data point is then appended to the fake dataset and the marginal likelihoods are calculated and averaged. In general, a single experience does not constitute adequate evidence for switching to an alternative model, since it lacks sufficient statistical power (Fig. 2.4). Note, that this claim is not true in extreme cases: there always exist outliers such

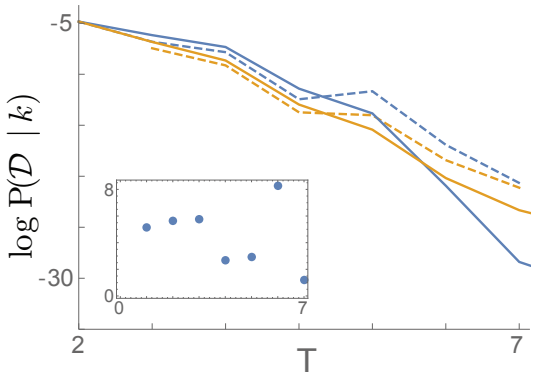


Figure 2.4: Inability of the memory-constrained learner to increase model complexity. Evolution of the true mLLH (analytic batch learner) of different models (*continuous lines*) and the mLLH of a constrained semantic learner (*dashed lines*) which only stores the (MAP estimate of) mixture components for earlier data points. A single data point is insufficient to induce model switch between $k = 2$ and $k = 3$ models (blue and orange, respectively). Inset shows an example data set that is selected for demonstration purposes.

that the marginal likelihood's automatic Occam's Razor effect will be overpowered by the unlikeliness of the new data (data not shown). If the present model estimate is correct, and the observed data corroborates this model then it can be integrated without information loss. We argue however, that models of differing form and complexity have different kinds of regularities that they can capture, and it is exactly the recurring appearance of features of the data that the current model is unable to represent that necessitates model change. Consequently, when a novel data point arrives which pushes the learner toward a change of model form but is insufficient in itself to force a switch, then the information loss prevents any subsequent model change (Fig. 2.4). This results in an inability to switch models for the memory-constrained learner even after observing arbitrary amount of evidence that supports a different one.

2.4 Episodic learner

The episodic learner differs from the semantic learner only in an additional limited capacity storage for observations. Since the semantic learner's inability to change models is a result of loss of information about past data, it is reasonable to expect that providing a buffer for data points is bound to help. However, we also require that the capacity of episodic memory necessary to enable model change should be small relative to the memory demands of a

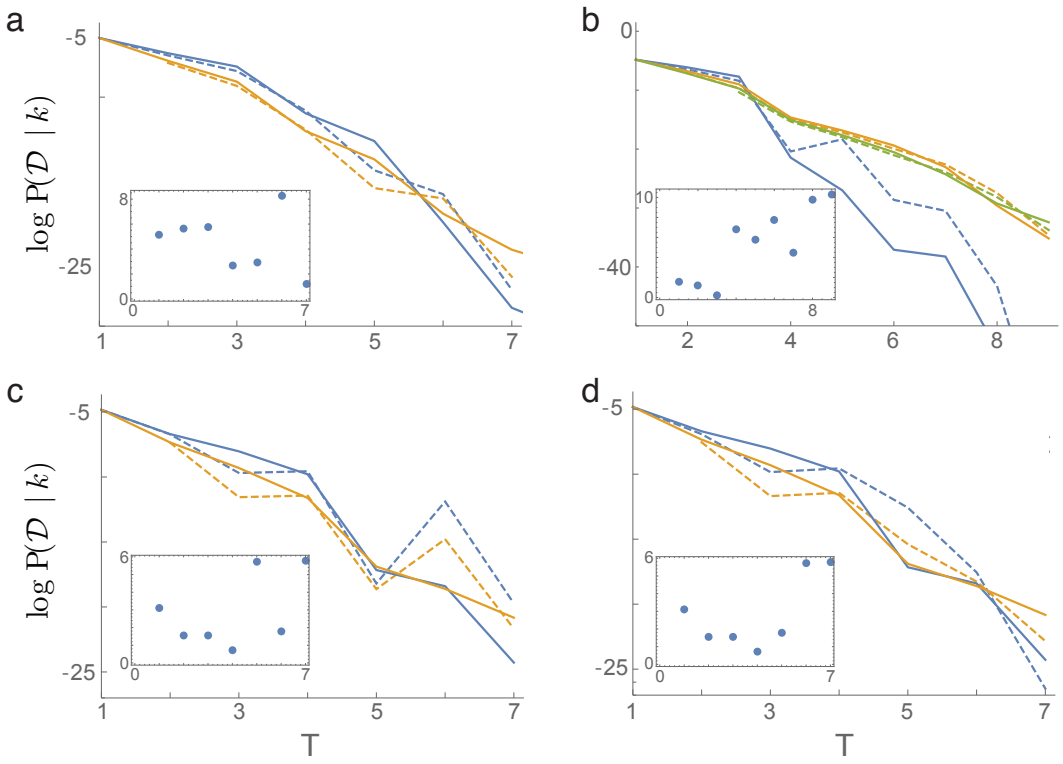


Figure 2.5: Ordered data and the effect of introducing episodic memory. mLLHs of the analytic batch learners (*solid lines*) approximate learners (*dashed lines*). **A, B,** Using ordered data and retaining the last two data points, the more complex model can obtain sufficient statistical power to overcome the Occam's razor effect at transitions $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$, respectively. **C,** For sampled data (unordered) a sliding window for two data points is insufficient to induce model switch. **D,** Episodic memory effectively rearranges data points (compare with panel **c**) such that the arrival of a subsequent data point(s) incompatible with the simple model induces model switch.

batch learner. Simply using this storage indiscriminately as a sliding window is inefficient for enabling model change (Fig. 2.5) since the experiences that taken together would provide the necessary statistical power for model change might not arrive consecutively. Taking full advantage of episodic storage requires the learner to optimise its contents and use it selectively. Thus, given a bounded capacity, the selection criterion for determining which data points to store in episodic and which in semantic memory is expected to be optimised to support the learner in dealing with online model selection.[1]

In order to retain statistics necessary for model transitions, an episodic learner needs to identify points that have a large information content with re-

[1]Note that here we make a binary choice between storing each observation in either episodic or semantic memory, in order to avoid double-counting the information in the observation.

spect to fitting the models. The Shannon definition of surprise, $-\log(P(\mathcal{D}|m))$, has been criticised as being unfit for this purpose because low predictive probably does not guarantee that the observation is informative with respect to the appropriateness of the model. Therefore we adopt the Bayesian definition of surprise [72], which characterises the extent to which the posterior is different from the prior expectations

$$S(\mathcal{D}, m) = \mathrm{KL}(P(m|\mathcal{D})||P(m)). \tag{2.10}$$

Ideally, episodes that are maximally informative regarding the model form would be sought but that would require evaluating the model posterior, $P(m|\mathcal{D}_{T-1})$, which is not accessible, since the learner doesn't necessarily evaluate the same set of models at different steps. Instead, we use the surprise in the model parameters as a proxy: this selects observations that change the learner's beliefs about the parameters the most. A large change in the parameter posterior signifies a difficulty in explaining the new observations and previously seen data under the current model which suggests that a change of models might be appropriate. Another insight can shed further light on the motivation behind our choice of selection criterion. Adopting the perspective that the memory trace is a lossily compressed form of the data, it should be optimised so that the distribution over past data – used in approximating the mLLH and alternative posteriors – is going to be as accurate as possible. We can view the combination of episodic and semantic memories as jointly providing a representation of the agent's past experiences $P_{SM}(x|\eta) + \sum_{x_m \in EM} \delta(x - x_m)$. In order to achieve the best compression the learner needs to use each kind of memory system to store the information it is most suited to reconstruct. Performing such an optimisation would be relatively straightforward by comparing the combined representation with the data, but the data was previously discarded. The learner can, however, select the data points that would change the reconstruction to a large extent, by seeing how much the posterior would change if the given experience was stored in semantic memory. Taken together, we formulated the criterion for selecting a data point for storing in episodic memory by assessing whether the dissimilarity of posteriors with the novel data exceeds a fixed threshold:

$$KL(P(\mu|x_T, \eta, k)||P(\mu|\eta, k)) > \tau. \tag{2.11}$$

Threshold τ is measured in units of surprise and its value was determined empirically, but performance is relatively robust to its choice. At low threshold values the learner becomes non-selective, which results in accumulating sequential mini-batches. On the other hand, at high threshold levels the learner will be reluctant to store anything in episodic memory and is thus asymptotically equivalent to the semantic learner. When episodic memory is saturated the learner "consolidates" the episodes by performing batch learning on its content. Upon triggering a model change the episodes also serve to find the parameter posterior of the novel model. For demonstration we have set the maximal size of episodic memory to one and used it to show that the problems of a constrained semantic learner can be effectively alleviated (Fig.2.5).

We have directly contrasted the performance of learning models in the model selection task on random data sets of length $T = 12$ where the generating distribution had $k = 2$ or $k = 3$ components (Fig. 2.6). Besides the unconstrained learner and the semantic learner, we set up models for an episodic learner with a memory capacity of one and two items, and also a pseudo-episodic learner that does not perform optimisation on the items to be stored in episodic memory. The episodic learner can demonstrate a remarkable increase in performance even with an extremely limited capacity. In order to make a fair comparison between $k = 1 \to 2$ and $k = 2 \to 3$ switches we balanced the difficulty of model switch through the variances of the Gaussians. Our analysis on $k = 2 \to 3$ switch revealed an even more pronounced advantage of the episodic learner over the semantic learner, doubling the probability of a correct switch.

2.5 Order effects in a toy model of Flesch et al.'s tree planting task

As we have seen, the unconstrained learner is insensitive to the order of data points, while the constrained learners are not. Specifically, Fig. 2.5 shows that the semantic learner is better at switching to a more complex model when

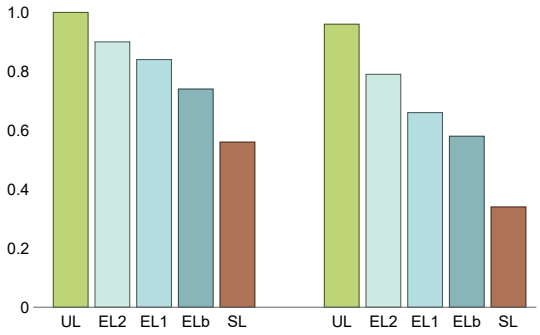


Figure 2.6: Comparison of model learning in different learners. UL:unconstrained; EL2: episodic with capacity 2; EL1: episodic with capacity 1; ELb: pseudo episodic with no selectivity; SL: semantic. Probability of $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$ model switch when data comes from a MoG with $k = 2$ and $k = 3$ (*left* and *right panels*, respectively) estimated from a thousand model runs each.

data points arrive in a *blocked* fashion, i.e. when they are ordered such that sequential points are likely to come from the same component in the MoG. Standard neural network models of learning show the opposite pattern of sensitivity to learning schedule, where they are able to achieve great performance on i.i.d. or *interleaved* data, whereas in the blocked setting they typically only maintain performance on the last task they were trained on, suffering from what is known in the machine learning literature as *catastrophic forgetting* [73]. Human learning typically exhibits the former pattern: In a particularly striking demonstration Flesch et al. [74] showed that humans, as opposed to standard neural networks, benefit from blocked training in task switching paradigm. In the following, we extend our episodic toy model to a simplified version of the Flesch et al. experiment and demonstrate that the effect shown in Fig.2.5 applies to this setting as well.

2.5.1 Experimental setting

In Flesch et al.'s experiment, participants were tasked with deciding which type of tree would thrive in two different gardens (the north and south garden). Trees varied in terms of branch and leaf density (Fig. 2.7A). Each trial began with an image of either the north or south garden as a contextual cue, followed by an image of a tree, which participants could choose to plant or reject. Based on these rewards participants could discover that only one of the tree

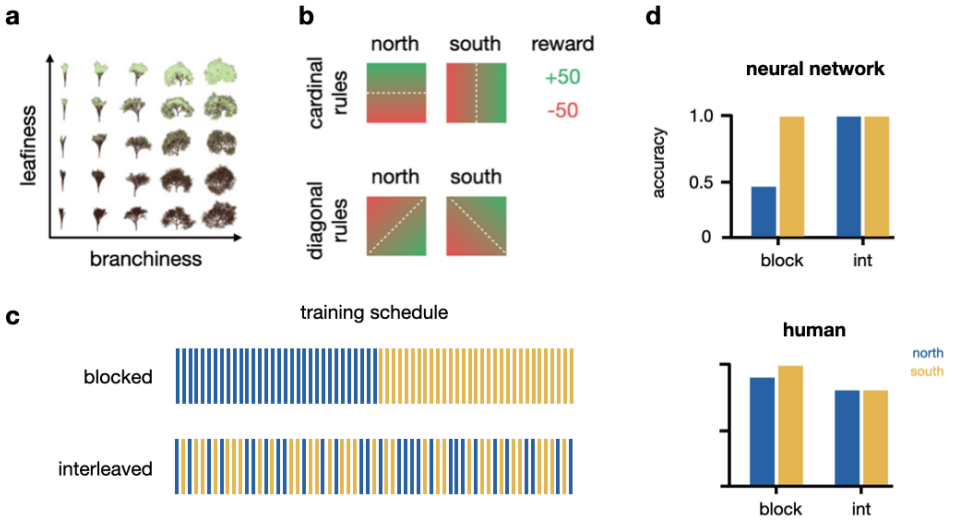


Figure 2.7: Task switching experiment of Flesch et al. [74]. **A,** trees varied along two latent feature dimensions, called 'branchiness' and 'leafiness'. **B,** the underlying rule for which trees grow well in each garden could be one of two types. In the 'cardinal' condition, a single latent feature determines reward in each garden, whereas in the 'diagonal condition', both features are required. **C,** trials were presented for each subject in one of two possible training schedules: under the 'blocked' schedule, subjects were first presented with all training stimuli from the first garden, followed by all stimuli from the second. Under the 'interleaved' schedule, the training stimuli were shuffled randomly. **D,** Task accuracies (probabilities of making the correct decision on planting each tree stimuli) in each garden at the end of training for a vanilla ANN (top) and average over human subjects (bottom).

features was relevant to how well trees would grow in each garden(Fig. 2.7B, cardinal rule). Participants were trained with either a blocked or an interleaved curriculum and were evaluated on an interleaved test block without feedback (Fig. 2.7C). Flesch et al. have found that humans, as opposed to standard neural networks trained with stochastic gradient descent, perform better in the blocked regime (Fig. 2.7D).

2.5.2 Computational model

In order to construct a minimal model of how the order effect arises in the episodic learner in the tree planting experiment, we explored a simplified setting where each observed variable is restricted to binary values: the garden or context c, the tree features z_i and the reward r. In this simplified setting each trial is one of 16 possible observations. The algorithm we used for the learners was identical to the one used in the MoG setting with the exception of the

following simplifications: i) instead of computing the parameter posteriors and mLLHs analytically, we used sequential importance sampling for approximating them and ii) instead of reconstructing the posteriors with the variational approximation, we used the posterior of the dreamed dataset with the highest mLLH on the winning model as the reconstruction. Note that this latter approximation for the posterior is much more noisy than the method used in the MoG setting, as it only takes into account a single dreamed dataset.

The hypothesis space in which structure learning operates is depicted on Fig 2.8. An important distinction between possible models in this hypothesis space, analogous to learning the value of k in the MoG setting, is whether they learn separate decision rules for the two gardens (2x_ models) or not (1x_ models). A second distinction is whether reward is dependent on either one or the other tree feature(_x1D), or both of these features(_x2D models). Our naming convention is based on specifying how many decision boundaries the models use, followed by the dimensionality of the decision boundaries in tree feature space. Each model had a discrete direction parameter (the direction of the decision boundary) with a uniform prior, and a continuous noise parameter with a beta prior. The noise parameter specifies the likelihood of observing a reward for a tree that should not be rewarded according to the decision rule. This parameter can incorporate both the possibility of noisy rewards given tree features, as well as uncertainty in the estimation of the tree features given the image of the tree for the subject. The 2x_ models are composed out of the 1x_ counterparts.

2.5.3 Results

In this hypothesis space, the automatic Occam's razor orders the model by increasing complexity. Due to the low dimensionality of the space of observations, we can enumerate all possible data sets consisting of a low number of trials (Fig. 2.9). Note however that interestingly, each model has the same mLLH on the first observation (not shown). We have generated data sets from a cardinal model (2x1D) without noise and with $T = 11$, as in this simplified setting this was typically sufficient for the unconstrained learner to discover the correct model structure. Computing the proportion of successes over 100

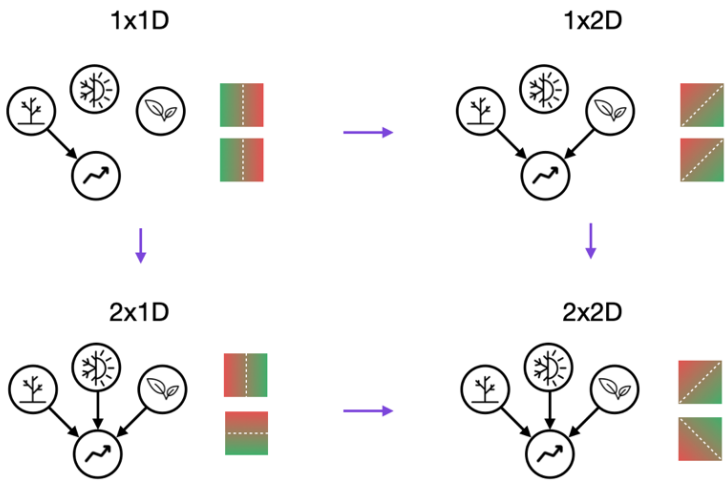


Figure 2.8: Space of possible model structures in the computational model of a simplified version of the tree planting task. The simplest candidate model where reward is determined in both tasks based solely on a single feature is shown on the top left. Model complexity can be increased in two ways: i) stepping to the right, the second tree feature is also taken into account in the prediction of reward, ii) stepping downwards, the garden is also taken into account, that is, the learner introduces a second decision boundary for reward prediction, which is the ground truth model structure in the cardinal condition. In the bottom right we see the model that uses two decision boundaries with two features each, which is the ground truth model in the diagonal condition.

runs each, we have found that similarly to the MoG model, episodic memory is essential in discovering the correct model structure (Fig 2.10A). Furthermore, by looking at the same measure but varying whether the same data points arrived in a blocked or an interleaved fashion, we have found a benefit in the blocked condition for the episodic learner (Fig. 2.10B) as well as all versions of the constrained learners (not shown).

2.6 Discussion

We have offered a normative argument for the existence of episodic memory by analysing a computational problem that the brain has to solve, namely online model selection in an open-ended model space. We used a simple minimal model to demonstrate that the introduction of memory constraints has dire consequences for a semantic-only learner and showed that these problems are substantially mitigated by an episodic memory, the contents of which are selected based on the Bayesian formalisation of surprise.

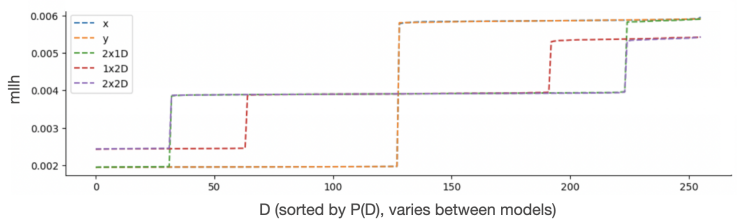


Figure 2.9: Marginal likelihoods of all possible datasets containing two observations. The automatic Occam's Razor effect orders models by complexity even with a flat prior over model structures. Cf. Fig. 1.14. Note that the horizontal axis is ordered by the mllh-s and differs between axes, therefore the fact that the two 1x1D models (x and y) overlap does not mean they are indistinguishable, only that they are of equal complexity.

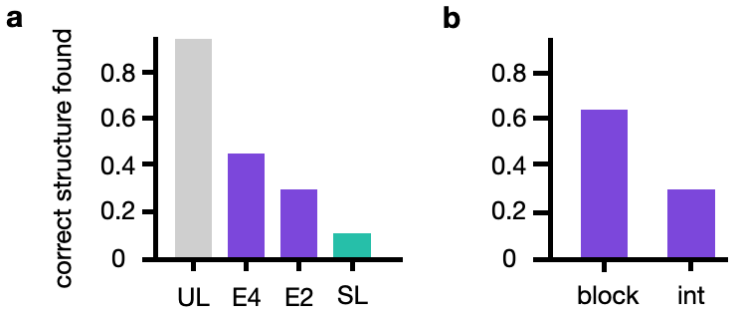


Figure 2.10: Results of toy model in the simplified tree planting setting. **A**, Proportion of 100 runs in which correct model structure (2x1D) was discovered for the various learners. UL: unconstrained learner, E4: episodic learner with episodic memory size of 4, E2: with size of 2, SL: semantic learner. Each learner was trained with an interleaved schedule. **B**, Effect of training schedule. int: same as EL2 on previous panel, blocked: same learner on same data set as int, but observations were reordered so that learner first observed all points from first garden then from second garden.

Our choice of model was motivated by analytical tractability which helped us to set a benchmark to model learning. While this choice constrained the form of the model and the size of the data set, the demonstrated problem is fundamental. These restrictions can be lifted by allowing the iterative posterior updates to be approximate, for example by using particle filters. Importantly, we strove to only use principles and approximations that are agnostic to the model class, so that the episodic learner can straightforwardly be extended to richer hypothesis spaces.

The overall goal of our normative account is to shed light on the dynamics underlying the organisation of long-term memory: from a continuous stream

of experience, how does the human brain determine what parts to remember and what to forget? It is extensively documented that humans are prone to systematic biases in these decisions. We share the widely-held belief that these systematic memory errors reflect rational adaptations to computational resource constraints. In our assessment, a comprehensive explanation and detailed predictions on how these processes work requires an understanding of both the computational function and the constraints that shape the dynamics of long-term memory. In this work, we aimed to provide the computational backbone for such a normative understanding: although some aspects of the current treatment are reminiscent of the characteristics of the dynamics of human memory (e.g. storing detailed representations of surprising events), a more direct comparison between model predictions and human performance will require the analysis of model classes that can be related to available human data.

Chapter 3

Semantic compression of episodic memories

It has long been known that human memory is far from an exact reinstatement of past sensory experience. In fact, memory has been found surprisingly poor for even very frequently encountered objects such as coins [75], traffic signs [76] or brand logos [77]. Rather than being random noise however, the distortions in recalled experience show robust and structured biases. A great number of experiments have shed light on systematic ways in which the distortions in recalled memories can be influenced both by past and future information, as well as the context of encoding and recall. Canonical examples of past knowledge influencing recall include experiments of Bartlett [78] where for folk tales recalled by subjects of non-matching cultural background, the recalled versions were found to be modified in ways that made the stories more consistent with the subjects' cultural background, leading to the suggestion that memory seems to be more reconstructive than reproductive. Various manipulations of the encoding context have been shown to influence the recalled memory; for example, presenting a label or theme before ambiguous sketch or text stimuli modulates both recall accuracy and the kinds of distortions that appear in the recalled memory [79, 80, 81]. Finally, paradigmatic examples of memory disruptions due to information obtained after the experience being recalled include post-event misinformation [82], imagination inflation [83], hindsight bias [84], or leading questions [85]. The rich set of systematic distortions provides insights into the principles governing memory formation, which

ultimately provides a means to predict how experiences are transformed in memory.

Traditionally, systematic biases in recalled memories have been interpreted as failures, confabulations of an unreliable memory system [86]. In contrast, adaptive accounts have been proposed that view these biases as being regrettable but necessary byproducts of adaptive processes in the brain such as generalisation, fast recognition or creativity [87, 88, 89]. Going even further, Bayesian accounts of reconstructive memory argued that in some cases, the memory distortions can be adaptive even if the goal is accurate recall, since previous knowledge can be used to correct for inaccuracies and fill in missing details in incomplete memories [90, 12, 91]. The Bayesian account provides a principled way of decoding memory traces by combining prior statistical knowledge with noise-corrupted information retained from the observation. However, it leaves a fundamental question open: in a normative model of memory what information needs to be retained and what pieces of information should be sacrificed to satisfy constraints on memory resources?

In this paper we argue that viewing the transformation of sensory experiences into memory traces as compression provides a normative framework for memory distortions. Specifically, we consider lossy compression, a form of compression where limited-capacity encoding is achieved at the price of imperfect decodability of original data, to characterise information loss during encoding. We point out that by adapting the mathematical framework for lossy compression, called rate distortion theory, to the constraints faced by the brain, we obtain semantic compression. Key to semantic compression is the assumption that a generative model of the environment is maintained in the brain. This generative model describes how the observed statistics of the environment has been generated from variables not directly observed through our senses [92]. According to semantic compression, it is the latent variables of this generative model of the environment that are used to compress experiences. If memory is optimised for natural observation statistics, then assessing the predictions of lossy compression regarding memory distortions requires generative models capable of handling such complex structured data. Constructing such generative models and performing inference in them can be challenging therefore we capitalise on recent advances in machine learning: we use variational

autoencoders to learn approximate generative models of structured data [93]. Importantly, a form of variational autoencoders, the beta-variational autoencoders can be viewed as a variational approximation to rate distortion theory, and we use this link between rate distortion theory and generative models to provide a theoretical framework for a unifying explanation of a large body of experimental data in the domain of memory distortions.

In the following, we introduce the theoretical background for semantic compression and then apply semantic compression to three different domains in memory distortions. First, we introduce basic concepts of rate distortion theory. Second, we introduce variational autoencoders and their relationship to lossy compression. Next, a paradigmatic example of memory distortion induced by past experience, domain expertise is investigated. In the coming section we discuss contextual effects through semantic compression. Finally, we discuss how gradual change of compression as time progresses incurs changes in memory traces.

3.1 Theoretical framework

3.1.1 Rate distortion theory

The branch of information theory that deals with data compression where information is lost during the process is rate distortion theory (RDT). According to RDT, a compact code is constructed that can be used to encode any data point in the dataset. A central insight of RDT is that there is no single optimal encoding: a trade-off emerges between the memory resources (rate, R) that are used for storing a given observation, i.e. the length of the code and the expected amount of distortion (D) in the recalled memory, i.e. the reconstruction of the original data from the stored code. Any given compression algorithm can be characterised by the trade-off it makes between these two quantities, and thus defines a point in the rate distortion plane (RD plane). An encoding, Q, can be improved by decreasing the expected distortion, D_Q, without increasing the rate, R_Q. Thus, the best encoding under any given memory resource constraint R is the one that minimises the expected

distortion when the rate is maximised at a specific value, R:

$$D(R) = \inf_{Q}(D_Q),\ s.t.\ R_Q \leq R,$$

or alternatively, by achieving a lower rate without increasing the expected distortion. Achieving the lowest possible distortion for each possible rate traces out the RD curve, establishing the range of possible optimal encoding schemes for a given distribution over observations.

Under any given encoding, the distortion of an individual observation ($d(x, \hat{x})$) between the observation (x) and its reconstruction ($\hat{x}$) can change across observations. The distortion term measures the expected amount of error that the encoding algorithm makes over the whole set of observations. The function $d(x, \hat{x})$ characterises how acceptable the distortion is and thereby defines an ordering across pairs of observations and their reconstructions, which defines the contribution of an observation to the the distortion term of RDT, $D_Q = \mathbb{E}_x[d(x, \hat{x})]$. An optimal lossy compression algorithm will selectively prioritise information such that alterations that are inconsequential according to this measure are discarded first.

An alternative formulation of finding the best distortion, D_Q, under the constraint of limited rate, R_Q can provide additional insights into the continuum of solutions obtained at different rates. The Lagrange-multiplier formalism is used for constrained optimisation of the distortion such that instead of minimising D_Q one needs to minimise

$$L = D_Q + \beta R_Q,$$

where the constraint of fixed rate when optimising the distortion is formulated through the Lagrange-multiplier β, which sets the trade-off between two terms. This formulation of the objective is only applicable when the RD curve is strictly convex but this requirement is often fulfilled in practical cases. According to this Lagrangian formulation, any compression method can be associated with a point on the RD plane, with optimal algorithms lying on the curve. Every point on it can be identified with a single value of β, which is the local slope of the curve. Thus, β directly corresponds to a particular point on the rate-distortion trade-off continuum: for example a high value of β is

associated with strong compression, yielding a low rate but high distortion.

While RDT provides a normative framework for lossy compression, the errors or compression artefacts resulting from traditional lossy compression algorithms of images or video (such as block boundary artefacts characteristic of JPEG) look qualitatively different from the errors committed by human memory [94, 95]. If one intends to use RDT as a normative framework of human memory errors then such mismatch ostensibly casts doubt on the applicability of the framework to memory phenomena. However, compression of sensory experience for the human brain is characterised by a number of constraints which distinguish it from the traditional problem of compression of image and video data. We propose to accommodate these constraints in the framework of RDT to obtain semantic compression.

3.1.2 Semantic compression

A fundamental difference between RDT and the form of compression required by the brain is that while the former produces optimal reconstructions given a known source distribution, the distribution of observations is not known for the brain but has to be learned from experience over time. Specifically, natural observation statistics define richly structured and high dimensional distributions, which have to be learned from a comparatively limited set of observations. In machine learning, this challenge is addressed by generative models. In order to cope with severely limited training data, generative models include inductive biases such as restrictions on hypothesis spaces, priors or hyperparameters. These biases influence decisions regarding what features of the data are generalisable to future observations and what features should be deemed random noise. We argue that since the problem of generalising from a small amount of observations to the true underlying distribution is a fundamental challenge in both generative models and compression in memory, similar inductive biases have to be incorporated in both. Therefore, we propose that the normative approach for adapting RDT to the problem of human memory is through compression via generative models.

Probabilistic generative models have been implicated in understanding human and animal behaviour in a multitude of cognitive tasks [96, 97, 98, 34].

For example, perception has been previously cast as a process of unconscious inference, which is aimed at inferring the latent state of the environment based on noisy sensory observations [92, 26]. This inference can be accomplished optimally by inverting the generative model which describes the way latent variables give rise to observations. Another domain where generative models have found support is action planning, where the model is used as an environment simulator to predict likely consequences of actions [30]. Following previous research, we assume that a statistical model of the environment is maintained in semantic memory and formalise semantic memory as a probabilistic generative latent variable model of the environment [10, 11, 54]. We argue that semantic memory represents the best estimate the brain has of environmental statistics, and therefore assume that it is this approximate model of the environmental statistics that compression is optimised for.

When the RDT framework is applied to the compression problem faced by the brain, a further issue needs to be considered: RDT does not specify how distortion should be measured, leaving it to be defined by the application. The distortion function represents the agent's judgements on how relevant particular features of the observation are and efficient compression hinges on selectively retaining this information. Thus, the definition of the distortion function raises the question of what parts of experience are relevant for the human brain. We argue that this problem is identical to that encountered in perception, where computations aim at extracting the latent variables underlying the activity of sensory neurons. In semantic compression the distortion function is defined by the generative model maintained by semantic memory, and the relevant features of observations are those that are extracted into the latent variables of this model. Importantly, such a choice for the distortion function and optimising for the complex structure in natural observation statistics can yield qualitatively different errors from those made by traditional compression algorithms, which only exploit simple, low-level regularities in image statistics.

3.1.3 Variational approach

Generative models and rate distortion theory are separate frameworks developed with largely different goals, however there has been a flurry of recent

work pointing out connections between the two [99, 100, 101]. A recent development in machine learning is the introduction of variational autoencoders, which can effectively learn a generative model of complex, high dimensional distributions as well as perform approximate inference over the latent variables. Interestingly, recent studies have established a link between a variant of variational autoencoders and rate distortion theory. Here we briefly describe the variational framework in which rate distortion theory and generative models can be jointly discussed.

Learning probabilistic latent variable generative models of natural stimuli such as images, videos and sound has been a major challenge in machine learning and requires approximate methods. Many of these models utilise latent variables, z, to factorise the distribution over observations, x. The set of latent variables can be thought of as the factors that contribute to the structure of the input data and constitute a representation of the data, often with lower dimensionality. These latent representations often show desirable qualities such as disentangling independent factors of variation. One of the most successful approaches to learning approximate latent variable generative models is a class of models called variational autoencoders (VAE) [93]. VAEs are capable of jointly learning the parameters of the generative model as well as performing inference by approximating the posterior distribution over latent state variables through variational methods. In variational Bayesian inference the true posterior distribution, $p(z|x)$, is approximated by a distribution $q_\phi(z|x)$ from a simpler distribution family parameterised by ϕ. Once such a distribution family is chosen, the goal is to minimise the dissimilarity between the true posterior and the approximate posterior:

$$\text{argmin}_\phi \, \text{KL}(q_\phi(z|x) \,||\, p(z|x)),$$

where KL is the Kullback-Leibler divergence, which quantifies the dissimilarity between two probability distributions. While this term cannot be computed directly, it can be shown that maximising the evidence lower bound (ELBO),

$$\mathcal{L}(\theta, \phi, x) = \mathbb{E}_{z\sim q_\phi(z|x)}(\log p_\theta(x|z)) - \text{KL}(q_\phi(z|x)||p_\theta(z)),$$

also minimises the KL divergence. In addition to learning to perform accurate

inference through optimising the parameters ϕ, the generative model is also learned through optimising the ELBO over the parameters θ. The generative model consists of the likelihood, $p_\theta(x|z)$, describing how the observed variables depend on the latents, and the prior distribution over latents $p_\theta(z)$. The first term of the ELBO is often called the reconstruction term, alluding to the fact that it penalises inaccurate reconstruction of the observation. The second term is usually viewed as a regularisation term, as it penalises deviations from a simple posterior. In VAEs the approximate posterior is typically of a simple form, such as a Gaussian, which is parameterised by its mean and covariance structure. VAEs also utilise amortised inference, where the computation of the approximate posterior is amortised by training a neural network to output the parameters of the approximate posterior on the training set. The output of this *encoder* then produces the parameters of the approximate posterior over z in later observations. An extension of VAEs, β-VAE, is particularly relevant for establishing a formal connection between generative models and RDT. β-VAE introduces a scalar multiplier, β, that scales the regularisation term and can effectively trade-off the two terms with the motivation that individual latent variables of the learned latent representation correspond to independent and interpretable sources of variation in the observed data, i.e. encouraging more disentangled representations [102].

To understand the relationship between β-VAE and RDT, we turn to a specific formulation of compression called the information bottleneck (IB) method [51]. The IB method extends RDT such that it guides the choice of the distortion function. The IB method introduces the term *relevant information*, the information that we intend to retain after compression. If the goal of compression is to lose information such that estimation of the relevant quantity, y, is minimally affected then we can formulate the relevant information as the information in the compressed representation z with respect to the relevant quantity y, that is the mutual information $I(z, y)$. Consequently, in the loss function of IB the goal of maximising relevant information is traded off with compressing observations through the latent representation, and the loss to be minimised becomes:

$$L_{IB} = -I(z, y) + \beta I(x, z)$$

It can be shown that minimising the IB loss function corresponds to a dis-

tortion measure that prioritises information that contributes to the prediction of relevant quantity y [51]. Although the IB method provides an algorithm for optimising the loss function, it is not feasible to apply to high dimensional naturalistic data. Alemi et al. [99] have shown that the IB objective can be efficiently approximated through variational methods. Importantly, the IB method formally defines a supervised objective since optimisation of the compression is achieved with the objective of optimising for a particular 'output' variable y, but an unsupervised version of the IB method can be constructed which gives rise to the same objective as the β-VAE. In this sense, β-VAE can be seen both as a generative model and a lossy compression algorithm: the reconstruction term in the ELBO can be interpreted as the distortion, D, and the information limiting regularisation term as the rate, R.

The correspondence between RDT and β-VAEs highlights the relation of latent variable models to lossy compression. Inferring a posterior over latent variables z upon the observation of stimulus x amounts to only retaining the statistics of stimuli captured in variations in z but discarding those beyond the sufficient statistics of the latent variables.

In summary, we use the framework of VAEs to learn the kind of generative model hypothesised to be maintained by the human brain and we link approximate inference over the latent variables of β-VAE to inferences made by humans. We then analyse this model from the point of view of lossy compression, allowing us to model and provide a normative explanation for a large variety of memory experiments.

3.2 Results

3.2.1 Domain expertise and congruency

According to semantic compression, efficient compression hinges upon accurate knowledge of environmental statistics. Since in the case of the brain these statistics are estimated based on experience collected over time, the accuracy of the estimate is expected to increase with the amount of experience within a cognitive domain. As the estimate becomes more accurate, compression becomes closer to optimal and consequently recall errors are expected

to decrease. However, this enhancement in recall accuracy is only expected to occur for observations congruent with the statistics of the domain, as a compression algorithm optimised for one distribution will be poorer at encoding observations coming from a different distribution. Assuming semantic compression, constructing artificial stimuli of the same domain but exhibiting statistical structure incongruent with that of earlier experience will increase recall errors.

Recall of chess board configurations

Chess is an ideal domain for computational analysis of expertise on memory performance due to a number of factors. i) The data is rich, possible configurations are astronomical; ii) chess games trace out a complex subspace of possible configurations; iii) 'natural' game statistics is well documented; iv) expertise is graded among individuals, allowing for a more fine-grained analysis of the relationship between expertise and recall performance. We capitalise on these properties of chess to test how expertise relates to memory performance in different conditions.

In a widely studied paradigm in memory research using chess [103], a chess board configuration is presented for less than 10 seconds, after which pieces are removed and subjects are required to reconstruct the observed configuration by placing the pieces on an empty board. Subjects are classified into four skill levels on the basis of their Elo points. Recall performance is measured in two conditions: In the case of '*game*' (or 'meaningful') configurations chess pieces are placed according to states taken from actual games, while in the case of '*random*' (or 'meaningless') configurations positions of chess pieces of game states were randomly shuffled.

We trained a β-VAE to learn the distribution of chess pieces during standard chess games downloaded from the FICS games database (chess-VAE). Briefly, a board configuration was represented as a 64 by 13 element matrix corresponding to the 64 positions and the 13 possible pieces, with an element of the matrix taking one if a particular chess piece appeared on a given position. This input was encoded with the β-VAE in a 64-dimensional latent space (for additional details please refer to the Materials and Methods section). In order to capture the varying amounts of experience that subjects have with

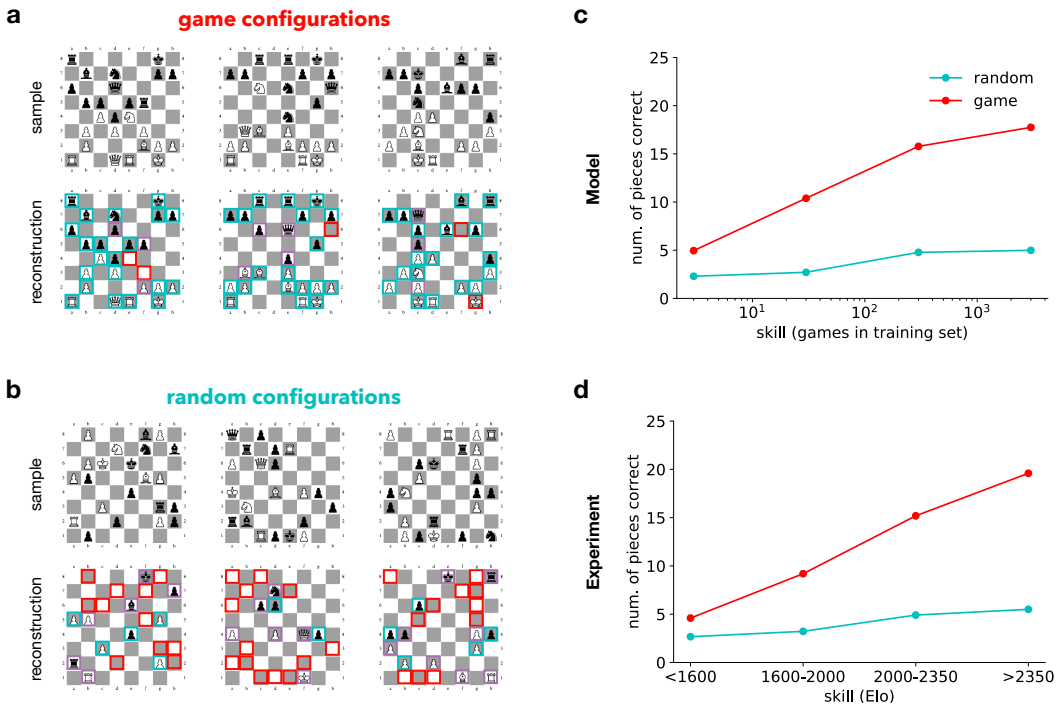


Figure 3.1: **Effect of domain expertise on memorising positions in chess. A,** *Top*, Chess board configurations from real game settings (*game configurations*). *Bottom*, reconstructions of the configurations from memory. These configurations are individual samples generated by the expert model based on the encoding of the presented configurations. *Green frames* indicate correctly reconstructed pieces, *red frames* indicate positions where a piece is missing or erroneously appears in the reconstructed game; *purple frames* indicate pieces whose identity is switched in the reconstruction. **B,** Same as A but instead of game configurations randomly shuffled pieces are presented (*random configurations*) and reconstructed. **C,** Reconstruction accuracy of the model for game and random configurations as a function of the training size. **D,** Reconstruction accuracy of human participants as a function of chess skill. Data reproduced from [103].

these statistics, we trained the generative model on varying amounts of chess games, using 0.1% (unskilled) to 90% (most skilled) of the entire training set consisting of approximately 250000 board configurations. In addition to the chess games, in order to mitigate overfitting to a low number of observations for the unskilled model, we augmented the training data with 10000 uniformly random board configurations at each skill level. This data augmentation can be seen as a hand-crafted inductive bias which optimises for a uniform distribution in the low data regime. Optimising for an uniform input distribution means that the algorithm maintains an ability to reconstruct any possible board configuration equally well, however since overall capacity is limited, this

means that no configurations can be reconstructed accurately. In the case of more skilled models the observations overwhelm the prior and consequently the prior has negligible effect. Note that we are taking a conservative approach in training the model, with no explicit instructions regarding the rules of chess or intent to win the game. Explicit knowledge of the rules makes certain configurations impossible or exceedingly unlikely, which can be utilised to aid recall. Nevertheless, reconstructions and unconditional samples show that the model captures an approximate version of these rules. To model the experimental recall setting, we used the inference network of the learned generative model to encode either game or random boards into a latent representation. Then, conditioning on the stored latent state we used the generative model to decode the memory trace into a reconstruction of the chess board configuration (Fig. 3.1A,B). Reconstructions by the model show the monotonic increase in accuracy for 'game' boards as a function of increasing chess skill (Fig. 3.1C). A similar monotonic increase in recall performance was found in humans (Fig. 3.1D), where recall performance ranged from around five pieces for amateur players to near perfect reconstruction for grandmasters.

In the 'random' condition, artificial stimuli obtained by randomisation destroys a significant portion of the statistical structure present in the configuration. In the case of 'random' boards the chess-VAE displays more errors both in omitted pieces and in exchanged piece identities (Fig. 3.1B). As a function of expertise, a monotonic increase in accuracy can be observed but with a distinctly smaller slope than in 'game' positions (Fig. 3.1C). In contrast to the 'game' stimuli case, for these artificial stimuli the accuracy advantage of skilled human players also shrinks substantially (Fig. 3.1D), meaning that the accuracy advantage of skilled players originates in the statistical structure of the stimuli.

Naively, one might expect that it would be easier to recall boards with only a few pieces, with the underlying assumption that storing the location and identity of each additional piece requires additional memory resources. However, board configurations containing most of the pieces are usually from early in the game, with the configuration strongly constrained by the initial state, whereas boards containing only a few pieces are typically from late in the game less constrained by the starting game setting. Intuitively, differences

in compressibility can be understood to arise from the relatively short description length of an early game configuration where one only needs to define the movements of a few pieces relative to the initial state. In summary, increased expertise in the statistics of a particular stimulus set specifically contributes to the enhancement of recall performance, which can be explained by recruiting knowledge stored in semantic memory for efficient compression of data.

Recall of pseudo-words

We have demonstrated that in semantic compression, accuracy of reconstruction improves with greater knowledge of the statistical structure of a given domain, and is also influenced by the congruency of the stimulus with environmental statistics. In the chess example, this was a binary choice, as each stimuli was either congruent (game boards) or incongruent (random boards). However, congruency can be on a spectrum, and in such cases, reconstruction errors are expected to decrease proportionally with it. Such an effect was demonstrated by Baddeley et al. [104, 105] in the domain of natural text.

Baddeley studied memory performance in an experiment where participants were required to reproduce synthetic words a short delay. Their experiment investigated the frequency of reconstruction errors as a function of both word length (which naturally increases memory load), and the level of congruence with natural statistics. They used synthetic words whose statistics was systematically matched to the statistics of letters occurring in natural language, where subjects could be expected to have similarly high level of familiarity. The method for creating these synthetic words is as follows: Zeroth order words are made up of random letters drawn uniformly. First order words are sequences of letters with the same distribution as in English language, drawn independently. Second order words are created using the probability of letter doublets, meaning the next letter is predicted based on the previous one. Third order words are created using the probability of triplets, where two letters predict the third one, and so on for higher order statistics. They have found that the degree to which the synthetic word conformed to the statistical structure of English text was strongly correlated with recall accuracy (Fig. 3.2A).

To develop a model of English word statistics, we trained a β-VAE on words from natural texts (word-VAE), using an architecture similar to the chess-VAE

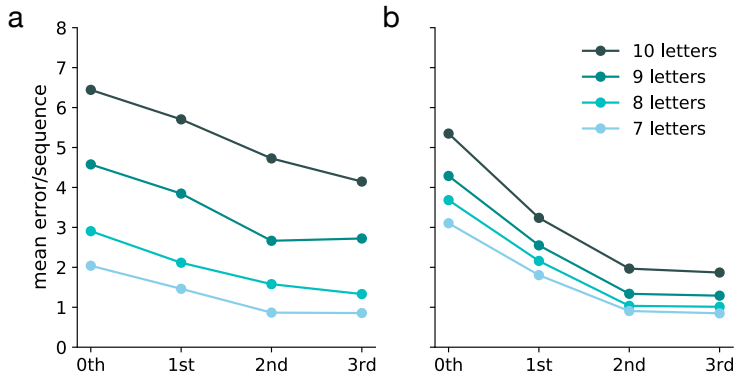


Figure 3.2: Effect of stimulus' congruence to natural statistics on recall accuracy in the synthetic word setting. Level of match between natural statistics and the letter-statistics of synthetic words was controlled by adjusting the level of statistics retained from natural words. *0th*: uniform letter distribution; *1st*: only frequency of letters is retained; *2nd*: frequency of letter pairs is retained, etc. **A**, Human performance for different synthetic word lengths, based on data from [105]. **B**, Performance of a β-VAE trained on natural text in a task analogous to the human experiment. Analysis of word-VAE and figure by Csenge Frater [57].

but with letters of the alphabet replacing chess pieces. By using the word-VAE to reconstruct synthetic words of varying lengths generated through the same method as the experiment, we observed a monotonic decrease in the number of errors with increasing order of approximation to natural text statistics for each word length (Fig. 3.2B). Notably, similar to the experimental data, recall accuracy curves are ordered by the number of letters, owing to the increasing entropy with word length.

3.2.2 Gist-based distortions

Semantic compression assumes that the statistics of stimuli is learned through a generative model and the latent variables of this generative model determine what features of experience are retained in lossy compression. Ideally, the latent representation that a generative model learns captures factors that explain a large amount of variance in earlier observations, are strongly predictive of future observations and rewards or allow for efficient manipulation of the environment. These latent variables are hypothesised to include lower level acoustic or visual features such as phonemes, or objects as well as abstract concepts such as what constitutes a good chess move or melody. These more abstract latent variables provide a high level, 'gist'-like description of the ex-

perience. By conditioning the generative model on the latent representation, observations that are consistent with the high level description can be generated. While precise details of the episodes will be lost during the encoding and decoding process, lost information can be supplemented by the generative model during decoding. More specifically, the generative model can be used to generate likely values of features for which the observed value was discarded during compression.

One consequence of reconstruction through a generative model is that memory will be sensitive to changes in the observations that affect the latent variables but allow for distortions that do not. At sufficient levels of compression this will result in falsely recognising or recalling items that were not themselves presented, but are conceptually related to items that were.

A second consequence of compression using latent variables of a generative model is that factors that influence the interpretation of the observation, that is the inference of values of latent variables, will also be reflected in the reconstruction. Specifically, in the case of ambiguous stimuli, contextual information influences the inferred latent representation, and consequently distorts the compressed memory by shaping lower-level details in ways that better conform to the shifted latent representation.

We demonstrate these effects in two experimental domains, the delayed recall of lists of words and recall of hand drawn sketches of objects. Note, that it is currently a challenge in machine learning to identify the computational principles that give rise to generative models that decompose observations into latent variables resembling the representations in human semantic memory. A particular advantage of using β-VAEs is that one of the main ingredients to achieve learning such a representation is thought to be the principle of encouraging disentangled features. β-VAEs have been shown to be able to discover disentangled latent representations from complex data in diverse domains and are therefore a good candidate for investigating the forms of memory distortions that result from the manipulations of latent representations in a domain-general way.

Intrusion of semantically related items during recall

One of the most extensively studied paradigms for reliably inducing strong false memories is the Deese–Roediger–McDermott (DRM) paradigm, where subjects have to recall lists of words. Language is a rich and computationally difficult domain, however recent successes in generative modelling of language suggest that the available size of corpuses makes it amenable to learning in an unsupervised way [106]. An inherent advantage of the DRM experimental setting is that since the recall order of word lists is not constrained, simpler, so called 'bag of words' models can be used which are significantly easier to train than text models that can also capture sequential dependencies between words.

In the DRM paradigm [107], subjects are presented with a list of semantically related words. The lists are created by collecting first associates of a particular common word, the lure word, from human subjects. In the experiment word lists are created from associates and presented to participants but the lure word is never shown during the memorisation phase. After a given delay that ranges from minutes to days, subjects have to either recall or recognise the studied words.

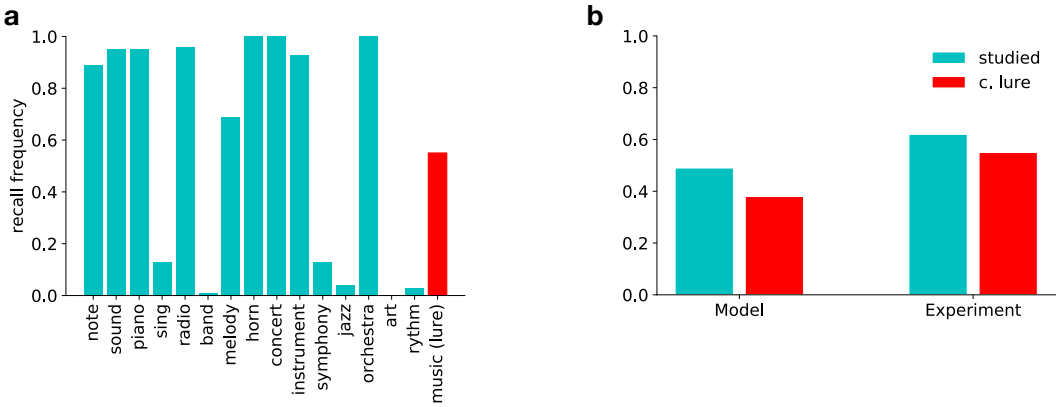


Figure 3.3: **Memory distortions for lists of words. A,** Frequency of recall from 100 samples for individual words in the text model trained on Wikipedia for the word list associated with 'music' from the DRM paper [107]. The lure word (*red*) is characterised by a recall probability comparable to the studied words (*green*). **B,** Comparison of recall probabilities for studied and lure words in the text-VAE model for 10 word lists (*left*, see Methods), and experiment (*right*) Roediger et al. [107].

In order to learn an approximate generative model for language, we have

used Wikipedia excerpts subsampled to 40 words for training a simple VAE architecture (text-VAE). The architecture was similar to the chess-VAE except that the noise model and input representation was adapted to text observations. This architecture has also been analysed previously in the machine learning literature as the Neural Variational Document Model (NVDM) [108]. The model takes text snippets as input to the β-VAE in bag of words representation, i.e. each occurrence of a word in a document is counted in a vector of dimension equal to the size of the entire vocabulary. This representation of the input disregards the sequential structure of text. The encoder mapping a document to a latent representation $\mathbf{z}$ of 100 dimensions consists of two dense layers of 2000 hidden units. The generative model is similarly structured and generates words independently (for additional details, see the Materials and Methods section). Wikipedia was chosen as a large and reasonably comprehensive corpus. After training the text-VAE model, for any presented list of words, a posterior can be inferred, which corresponds to the latent variables that might underlie the observed word list. Nearest associates of the lure words show that the learned representation captures similar statistical relationship between words as the methods used in the original experiments to generate associate word lists (25% of corresponding DRM list words appear in 50 closest associates to lure word in model, for examples see Methods). However, since the statistics of words in an encyclopedia is different from that encountered by a human during his lifetime, the model's interpretation of certain words can be biased (e.g 'chair' is strongly related to 'organization' in Wikipedia dataset but not in the DRM word lists).

Generative models of text such as topic models often make the assumption that natural text is concentrated on a low dimensional manifold in the space of all possible text data. If compression is optimised for reconstructing samples from a distribution with such manifold structure, then for observations that lie near the manifold the reconstruction will be drawn towards the manifold to an extent determined by available capacity. Furthermore, these generative models often contain an inductive bias, also characteristic of our text-VAE, that text is generated by independent latent factors of variation. Combined with natural text statistics this inductive bias results in the emergence of latent topics, which constitute clusters of semantically related words.

Any given text excerpt is represented as a specific mixture of possible topics. When such an algorithm is used to encode word lists from the DRM paradigm, the underlying implicit assumption of the model is that the list was generated by some mixture of a few latent topics. As capacity is decreased, the reconstructed observation will become an increasingly prototypical exemplar of the activated topics, leading to the intrusion of semantically related lure words. We used the trained model to test this hypothesis on the DRM paradigm. For this, word lists of the original paradigm were taken as the set of words coming from a document and we inferred a posterior representation associated with this document. This posterior was subsequently sampled and the synthesised word list was taken as the reconstruction of the original word list (Fig 3.3A). As expected, since semantically related words are likely to occur together in natural text, the model improved recall accuracy of such word lists relative to lists of randomly selected words, however the price is the intrusion of non-studied but semantically related words into the reconstructed list (Fig 3.3A). The intrusions indicate that while there is a loss of information, the encoding and decoding process keeps the reconstructed observation consistent with a stored gist level interpretation of the original observation. Importantly, the frequency of the recall of lure words is similar to the average frequency of the recall of studied words, reminiscent of human performance in the DRM task (Fig 3.3B).

Effect of varying contextual information on recall

We have argued that a second consequence of using the latent variables of a generative model for compression is that context influences both the degree and structure of distortions in recall. Hand drawn sketches of common objects have been used as complex naturalistic stimuli for exploring memory distortions, allowing the incorporation of contextual information by providing verbal labels or textual descriptions. The continuous nature of sketches allows us to explore graded and structured distortions of the observation along with the context dependence of encoding and recall. A dataset of millions of labelled sketch drawings created by a large and diverse set of human users of a browser-based game became available recently, and VAEs capable of handling these high dimensional data have been developed [109].

A well-known and robust example of the effect of contextual information is the experiment of Carmichael et al. [80]. In the classical experiment, intentionally ambiguous hand drawn sketches of objects from common categories were presented to subjects who were asked to reproduce these images after a delay. Two separate groups of participants were required to reproduce the sketches, with each group in one of two contexts. Context was established by providing a category name preceding the presentation of the drawings, with each name being consistent with one possible interpretation of the drawing. The authors found that, depending on the contextual cues, systematic biases were introduced in reproduced images that made the drawing more consistent with the provided label (Fig. 3.4B).

In order to analyse contextual effects in reconstruction in semantic compression we trained a VAE on sketch drawings from the QuickDraw data set (sketch-VAE). As an approximation of the semantic model for sketch drawings, we used the Sketch-RNN architecture, which models sketches not as raster images but as a series of sequential pen movements [109]. The model uses recurrent neural networks to make predictions on each subsequent stroke conditioned on its hidden state and the previous one and assumes Gaussian motor noise. We have selected ambiguous object-pairs from the QuickDraw data set and trained the model on 75000 drawings of each category. In order to model the effect of presenting a contextual cue, we have trained a conditional model on sketches belonging to each label. Each of these models represents a label conditional generative distribution $p(x|z, y = label_i)$ and a label conditional approximate posterior $q(z|x, y = label_i)$. During inference, we use the corresponding distributions to reconstruct the same ambiguous image. Consistent with human data presented (Fig. 3.4B), reconstructions from the conditional posterior resulted in systematic distortions of the original image consistent with the provided label (Fig. 3.4A). Systematic distortions introduced by the model were rich, spanning addition or deletion of features, rescaling of features, or subtle but characteristic changes in the shapes of reconstructed drawings (Fig. 3.4C). Systematic rescaling of features has been observed in humans [110], which is qualitatively similar to the rescaling found in the sketch-VAE model (Fig. 3.4D).

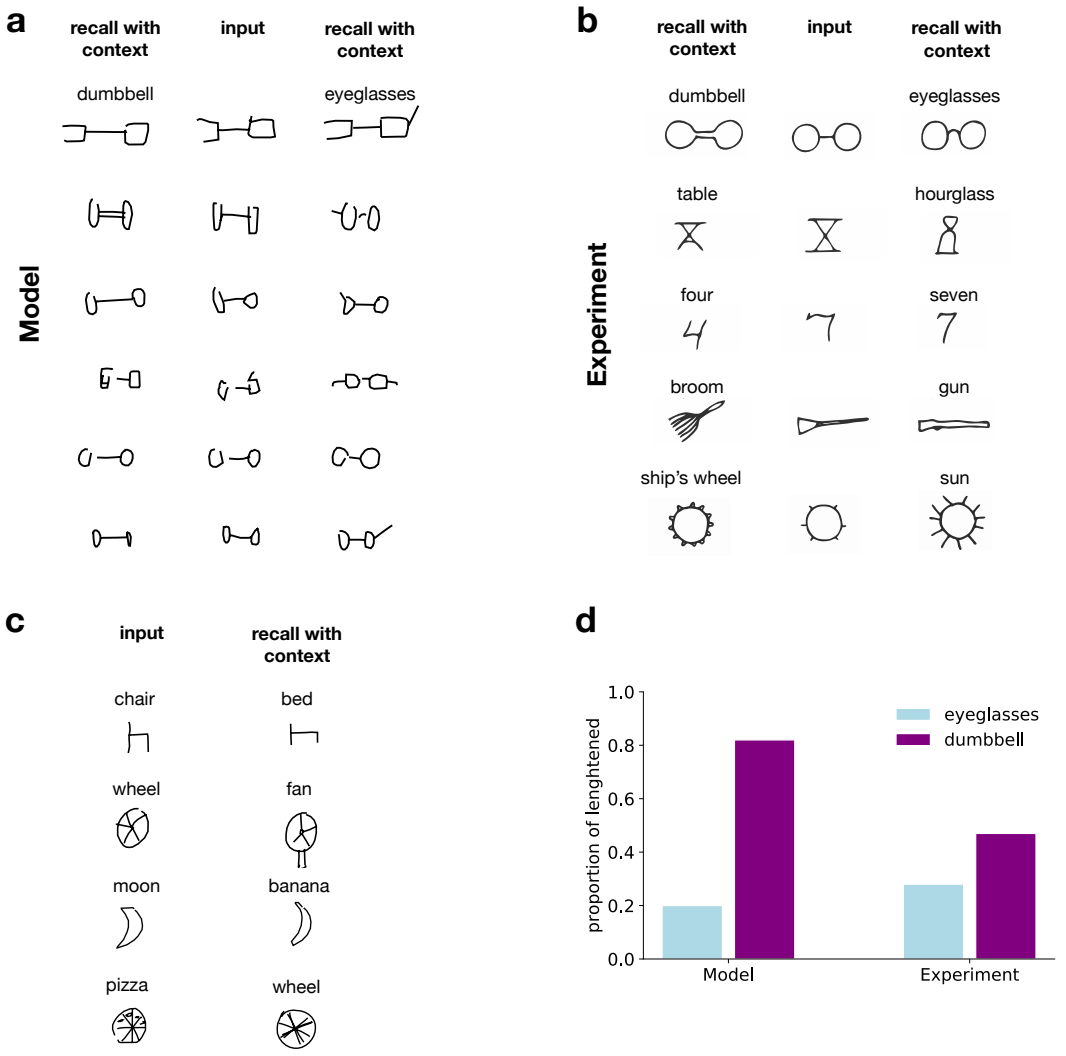


Figure 3.4: **Context effects on reconstruction of line drawing from memory. A,** *Middle column*, Ambiguous line drawings from the QuickDraw data set of eyeglasses and dumbbells. *Left* and *right columns*, Reconstructions of the image from memory in the dumbbell and eyeglasses contexts, respectively. Context is modelled by using a sketch-VAE trained on sketches from a single category with $beta = 2$. **B,** Examples of ambiguous drawings (*middle column*) and their reconstructions (*side columns*) when cues are provided to participants (*shown as text labels*). Data is reproduced from [80]. **C,** Effect of contextual information on the visual features in recalled stimuli in the model. Quantitative changes (*top*), qualitative changes (*middle*), and subtle changes in characteristics (*bottom*) occur as a result of contextual recall. **D,** Quantitative changes in visual features with changing context (proportion of the length of the line connecting circular features in the eyeglasses and dumbbell contexts) in the Sketch-RNN model (*left*) and experiment (*right*). Experimental data reproduced from [110]

3.2.3 Rate distortion trade-off

The value of retaining information from a given episode is likely to vary with respect to a multitude of factors such as how surprising the episode is, its relevance for predicting the near future or its emotional valence. As a consequence, we propose that memory resources allocated to storing episodes are unlikely to be constant either at the time of encoding or as a function of time. If memory resources are to be distributed rationally, this memory decay should not result in random forgetting as information theory provides a principled way of discarding information so that memories degrade gracefully.

Formally, optimal forgetting entails moving along the line of optimal encodings in the rate distortion plane in the direction of decreasing rate (Fig. 3.5A). At one extreme, where the rate distortion function intercepts the rate axis, resources are sufficient for lossless compression, meaning that verbatim recall is possible. At the other intercept, no information is retained relating to the individual episode and reconstruction is based purely on knowledge of environmental statistics. Starting from the point corresponding to verbatim compression, the memory trace becomes increasingly gist-like, until a point where even a very high level gist of the episode is lost. This way, the trade-off between rate and distortion results in the emergence of a continuum between gist and verbatim representations.

Temporal reduction of memory resources

In cognitive processes, such as prediction, time delay between storage of information and its retrieval is a factor that fundamentally affects its relevance and therefore the resources that should be dedicated to the particular piece of information. Anderson et al. [86, 111] proposed that it is a property of the natural environment that there is a decreasing need for information contained in individual traces as time progresses. For example, information contained in an email is more likely to be needed within a day of receiving a message than after a month. By studying library borrowings, access times of digital files, email sources and word appearances in the headlines of newspaper articles they concluded that forgetting curves demonstrate that human memory is adapted to this decreasing demand. These results have been corroborated by

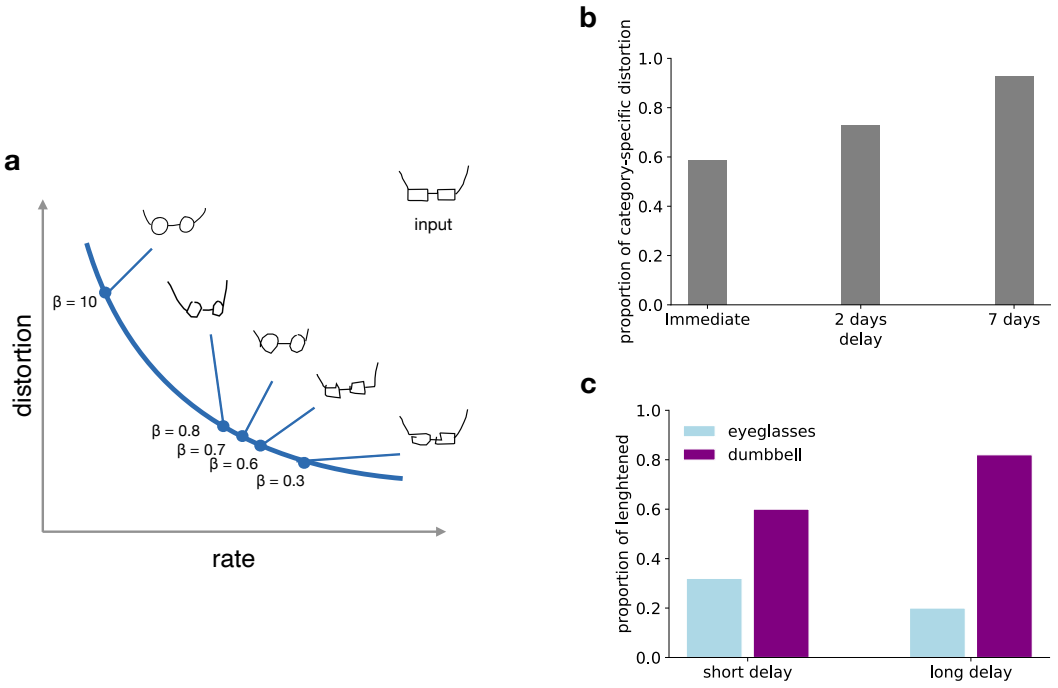


Figure 3.5: **Rate distortion trade-off in memory for sketch drawings. A,** Illustration of stimulus reconstructions as changes in β result in different points on the rate distortion curve for the sketch-VAE model. Inset image is used as input and is reconstructed with various levels of compression. Optimal forgetting implies moving along the curve in the direction of increasing β corresponding to increasingly prototypical reconstructions of the original drawing. **B,** Proportion of recalled sketches judged to show category specific distortions in humans due to the context presented during learning at different delays between stimulus presentation and recall. Distortions were evaluated by two of the experimenters and one judge naive to the purpose of the experiment. Data reproduced from [110]. **C,** Proportion of sketches reconstructed by the model showing category specific distortion as a function of increasing compression. Quantitative changes in visual features are assessed, similar to Fig. 3.4D.

forgetting curves of US presidents in multiple generations of college students [112]. This argument motivates the idea that different rate-distortion trade-offs can be studied by controlling the time between stimulus presentation and recall or recognition. The effect of retention interval has been studied both in the recall of hand-drawn sketches and even more extensively in the DRM literature, allowing us to contrast it with our model's predictions on targeting various points of the rate distortion trade-off.

In order to model the effect of delay, we have optimised models for increasing levels of compression by training them with increasingly larger βs. Since stronger compression implies more gist-like reconstructions, the high level con-

text has a stronger effect on the recalled drawing as memory resources are decreased in the sketch-VAE model (Fig. 3.5A). We assessed the scaling of features for the ambiguous eyeglasses-dumbbell stimulus in the two contexts as a function of available memory resources. The analysis demonstrated more frequent category-related distortions for βs associated with longer delays. (Fig. 3.5C). Similarly, higher levels of context related distortions were found for the same pair when recollection was tested with increasing amounts of delay with human participants [110] (Fig. 3.5B).

Several studies have examined the effect of delay on recall performance in the DRM paradigm [113, 114, 115]. Toglia et al. [113] has performed the experiment with recall immediately after presentation of the lists or after delays of one or three weeks. In contrast with most of the prior work on the subject, retention intervals were varied between subjects, avoiding artefacts due to retesting the same subject. They have found that while recall for studied words had fallen sharply, recall of lure words was relatively unaffected even after three weeks. In a variation of the original paradigm, for some of the subjects they have presented lists in a 'random' condition, pooling the words from six lists and presenting them in shuffled order. Interestingly, in the random condition recall probability of lure words had increased as compared to shorter delays at week 3. An even longer delay of two months has been studied by Seamon et al. [114], where they have found that eventually the recall of lure words also approaches zero. Thapar & McDermott [115] have looked at a similar design with a maximum delay of one week while also modulating depth of processing at the time of encoding. Many other studies besides these have examined the effect of various manipulations on accuracy for delayed recall but it has been a robust finding that memory for lure words, that is false memories, are more persistent in time than memory for studied words (Fig 3.6D).

We have investigated the dependence of recall on memory constraints in the word list recall paradigm using our text-VAE model. We have trained separate models for each memory constraint, corresponding to various levels of the rate distortion trade-off parameter β. We then used these models to reconstruct each word list at each setting of the memory constraint and subsequently evaluate recall accuracy for studied and lure words separately. Reflecting a transition from a verbatim-like to a more gist-like latent representation, recall

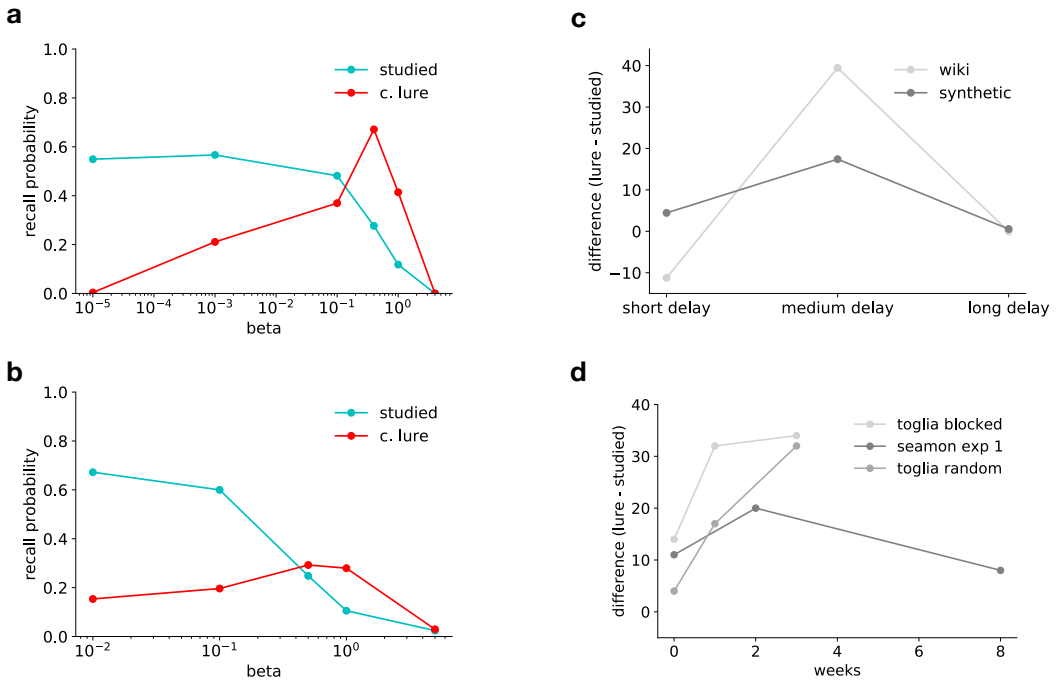


Figure 3.6: **Rate distortion trade-off in forgetting over time. A,B:** Using the text-VAE model, we modelled the dependence of memory representations on time by gradually increasing the compression rate, β. Recall probabilities averaged over multiple word lists of studied and lure words was measured on a text-VAE trained on Wikipedia entries (*top*) and on a synthetic vocabulary (*bottom*). Increasing the compression rate results in a monotonically decreasing recall performance for studied words. In contrast, increased delay of recall leads to an increase in false memories. For critical NS (lure) words the recall probability initially increases with larger compression rates but very high compression rates result in losing gist-like recall as well. Asymptotically the performance on semantically related S words will approach the performance on random word lists as less and less of the structure of the data is used. **C,D:** Difference between recall probabilities for lure words and studied words as a function of the delay between recall and study for the model (*top*) and experiments (*bottom*). Both Wikipedia-trained and synthetic vocabulary trained models predict persistence of false recall of non-studied lure words as compared to studied words, visible as an increase in the difference between lure and studied word recall rates as a function of time. For even longer delays, the gist information is progressively forgotten as well and consequently recall rates for both lure and studied words approach zero. The same pattern of increasing rate up to a delay of three weeks and a subsequent decrease can be observed in experimental data. Data is reproduced from Toglia et al. [113], Seamon et al. [114] and Thapar & McDermott [115].

performance on studied words decreased monotonically with increased level of compression. The decrease in accurate recall was paralleled by an increase in gist-based intrusions of related lure words starting from a negligible level as

the latent representation incrementally came to resemble a topic model. This was followed by a reversion of false recall rates to negligible levels, reflecting a loss of even gist information due to extremely limited capacity (Fig. 3.6A). Note that we have restricted our analysis to the main effect of length of delay in the delayed recall DRM experiments and thus our text-VAE does not explain the effects of manipulating depth of processing or word order effects. In order to control for the mismatch between Wikipedia and natural text statistics, in addition to the text-VAE trained on Wikipedia articles, we also tested model performance on synthetic data. We generated a synthetic corpus with controlled statistics from a Latent Dirichlet Allocation (LDA) topic model, which models statistical structure in text by assuming the presence of latent topics that are present in each document. In order to construct DRM lists for the synthetic model, we have used nearest neighbours according to word embeddings learned from the synthetic corpus (see Methods for further details). The synthetic data trained text-VAE shows a transient increase in lure recall, similar to the one observed with the Wikipedia-trained model (Fig. 3.6B).

We have argued that a decreasing need for information contained in a memory trace over time implies that memory resources assigned to the trace should be decreased. Therefore, according to RDT, the delay between encoding and recall corresponds to a change in the RD trade-off, controlled by the β parameter. However, the exact mapping between length of delay and the numerical value of β can't be derived from the theory but has to be calibrated based on measurements. In the standard DRM experiments, recall probabilities of studied and critical non-studied words are similar (Fig 3.3D), and consequently we set β corresponding to immediate recall to be the same as in our model of the original experiment ($\beta = 0.1$ for the Wikipedia trained model and $\beta = 0.5$ for the model trained on synthetic data). Increasing the rate from this level (decreasing β) results in increasing accuracy and thus the gradual disappearance of false memories. Since time delays are necessarily positive, the predictions of the model in this regime describe a hypothetical situation where memory resources available to the subjects were increased above what is typically measured in this paradigm. On the other hand, modelling memory decay in time by decreasing the rate results in a decrease in the recall of studied words, and an initial increase (for the models trained on Wikipedia) or much less pronounced

decrease (for synthetic data) in false memories, resulting in a similar pattern of relative advantage of lure words over studied words after medium-length delays as the one seen in human experiments (Fig 3.6C). At very high levels of compression (high β), recall for all word types is poor, as even the broad theme of the list becomes forgotten and the model samples lists of related words that occur together frequently in natural text.

3.3 Discussion

In this paper we demonstrated that the principle of lossy compression provides the basis for a unifying normative account of a wide variety of systematic memory errors. Central to our approach was that we related inference in probabilistic generative models to rate distortion theory: inferring the latent variables underlying observations amounted to selectively discarding information not represented in the latent variables, thus establishing a lossy compression method. We used a recently developed machine learning framework to train a probabilistic generative model on a variety of naturalistic, high dimensional data sets (chess games, line drawings, and natural text). We showed that the effect of domain expertise on recall accuracy in remembering chess board positions and gist-based distortions in remembering semantically related word lists arise as straightforward consequences of optimising the rate distortion objective. Furthermore, we demonstrated the emergence of varying degrees of 'gistness', resulting from the rate distortion trade-off.

3.3.1 Interpreting memory distortions as lossy compression

The experiments discussed in the study demonstrate key consequences of semantic compression in a single computational framework. The primary appeal of this framework is that it can integrate a large variety of experimental observations under a simple computational principle. The demonstrated effects, however, are more general than the experiments discussed here. For instance, the effect of domain expertise has widespread support in the memory literature not just in the domain of chess but also memory for sports trivia,

software code, medical images, and other games [116]. In addition to the effect of varying levels of expertise on recall accuracy, a similar effect arises if the congruence of stimuli to the statistical structure of the domain is varied along a spectrum, for example the order to which letter statistics of words conform to that of the English language [105]. The encoding of the observation into a posterior over latent variables can be understood as compressing sensory experience into sufficient statistics for the latents, which in addition to the gist-based distortion experiments analysed here explains seemingly paradoxical results that discrimination performance of sound textures decreases with increasing stimulus duration when stimulus samples come from the same texture family [117]. Such a process also implies that the level of difficulty of inference affects the accuracy of the recalled memory trace. Similar to the Carmichael effect, classical memory experiments have shown that providing a concise context which aids the interpretation of otherwise strongly ambiguous stimuli can greatly increase retention accuracy [79, 81]. Beyond the effect of expertise on reconstruction accuracy, the Chase and Simon experiments (1973) also display effects that are related to representing latent variables. In particular, temporal dependencies in placements of chess pieces were suggested to reflect chunking mechanisms. The proposed variational autoencoder framework naturally generalises to these domains as well.

The delayed DRM paradigm has been extensively studied [113, 114, 115]. In this study we only have addressed the effect of delay alone but effects such as divided attention, order effects and depth of processing were not discussed. Some of these could be addressed by natural extensions of this model, for example order effects would require extending the model to non-iid observations. Others, such as the effect of varying depth of processing during encoding or recall we do not see as direct corollaries of the RD framework and therefore would require further assumptions.

We have modelled forgetting as the effect of decreasing capacity allocated to a memory trace over time by training separate models for different levels of compression. A limitation of this approach is that there is no guarantee that a slightly compressed representation taken from a model with low β can be converted into the strongly compressed representation of the model given by a high β, as it is possible that the high β representation utilises information

that is present in the original observation but not in the low β representation (although in toy settings this seems not to be the case as the introduction of further capacity leads to capturing additional data generative factors [118]). Furthermore, taken as a process model this approach would require storing a separate semantic model for each available level of compression. Approaches where a single model is trained for compression at multiple rates have been proposed in the machine learning literature [119, 120]. In addition, hierarchical generative models have been introduced that are capable of generating observations at progressively increasing level of detail by conditioning on variables at more and more levels of the hierarchy [121, 122, 123]. Such hierarchical generative models define a straightforward process for converting a memory trace into one requiring lower capacity by selectively discarding information. The lowest level of compression is achieved by storing all levels of latent variables, then as time progresses, the states of successive levels of variables are discarded beginning with the one closest to the input layer. Some of these models show semantically meaningful partition of information between the layers in limited domains. For example in Maaloe et al. (2019), the compression process outlined above applied to portraits initially retains information about wearing glasses but discarding the specific information that those are sunglasses, and at later stages of compression it forgets about the glasses while keeping a large portion of facial features still intact. One way in which such a compression could be implemented is if semantic compression utilises the hierarchical representations in sensory cortices as have earlier been argued in [124, 125, 58].

3.3.2 Theoretical considerations

Application of RDT to lossy compression in the brain seems to be a natural choice. A closer inspection of the problem, however, reveals that from the perspective of compression and specifically its formal theory, RDT, a fundamental challenge arises. In memory systems, the data set used to learn the model is inherently incomplete, that is only a subset of the data that specifies the model had been observed. This setting defies a critical assumption of RDT that the data statistics are known. The brain tackles the issue of incomplete data by updating the model continuously when new data is observed. However,

the constraint that the statistics of observations is being learned concurrently with using it for compression places unique demands on a memory system. We have previously argued that if the overall structure of the model describing the environment is known then the parameters of the model can be updated once new observations are made and the only information that needs to be maintained is the sufficient statistics of model parameters. Consequently, other features that are not part of the sufficient statistics can be discarded without harm. However, if there is uncertainty over model structure and it is not *a priori* known which features are relevant for the model, then the ability to reconstruct the data becomes critical [54]. The need for such reconstructive ability motivates our use of unsupervised learning, which attempts to capture all of the variance in the data when resources are not bounded, constituting a perfect episodic memory. In case that relevance has to be sensitive to predictive ability [6, 126], rewards or a supervision signal, the framework can be straightforwardly extended through the same deep variational information bottleneck objective [9, 99]. Task variables can potentially also be accommodated in a generative perspective in which task variables are part of a generative model. Such models have been introduced in machine learning [127].

The variational approximation that we have used here provides a useful tool for integrating principles of RDT and probabilistic generative models, which can be tested under conditions where data complexity is close to that of the natural environment. Since this is the data set that human memory systems are adapted for, we believe that these are relevant stimuli to contrast capacity constraints of the model and that of human memory systems. Application of the framework, however, also comes with specific choices and alternative formulations are possible. In RDT and the IB method, the rate term is defined as the mutual information between the observation and the latent code which the variational method provides an upper bound on. This is an abstract constraint, and the specific influences on the latent representation depend on model architecture. For example changing the Gaussian prior to a Laplace distribution would result in a constraint on the sparsity of latent activations. Furthermore, it has been argued that memory resource constraints for the brain would be better captured by restricting the representational cost of storing the encoding, corresponding to minimising the entropy of the code instead of the mutual in-

formation[128]. This choice leads to the Deterministic Information Bottleneck (DIB) method, which can lead to qualitative differences such as the optimal reconstructions being deterministic. Variational approximations also exist for the DIB, and it has been argued that a discrete latent space variant of variational autoencoders called Vector Quantised Variational Autoencoder can be viewed as an approximation to the variational DIB principle. Furthermore, the correspondence between RDT and generative models can be drawn in alternative ways: Balle et al. [101] show a correspondence where the posterior is of fixed variance and the multiplication factor beta arises from the variance of the observation noise. We see the information theoretical form of the bottleneck constraint as an approximation to multiple constraint terms, possibly arising from the demands of cognitive functions other than memory, each having a contribution to shaping the latent representation. The question of which variant of the computational framework and what combination of constraints would correspond most closely to representations and memory distortions measured in human experiments is a subject for further investigation.

3.3.3 Related work

RDT has recently been proposed as a framework to investigate distortions of memory by Bates and Jacobs (2020) [129]. While the computational principles they apply and those in this and our previous work [130] have strong parallels, the differences highlight different aspects of using generative models for compressing complex data. While VAEs constitute state-of-the-art in machine learning for learning generative models of high dimensional data, even these models struggle to capture the full richness of natural stimuli. In particular, learning highly structured noise models has proven difficult, leading to issues such as blurred reconstructions [131]. In order to mitigate this problem, instead of using pixel image data, we have opted to use data represented in low level features such as chess board locations or pen stroke endpoints. This choice essentially circumvents the problem of learning the low level noise model for the network. Bates and Jacobs [129] take the alternative approach of working directly with pixel data and thus provide an end-to-end learned model. The choice of training the model end-to-end versus learning over low level abstract

features has complementary benefits: while using pixel data to study memory effects is certainly appealing since the perceptual process is more completely integrated, it has the disadvantage that the generative model needs to cope with limitations of VAEs on natural images, such as blurry reconstructions. Blau et al [132] proposed that the issue of perceptual quality of reconstructions could be mitigated by introducing an additional term to the RD trade-off, which could be optimised utilising the framework of the other prominent form of deep generative models besides VAEs, Generative Adversarial Networks [133].

We have argued that while we rely on the unsupervised version of the information bottleneck to make the connection between RDT and latent variable models, the information bottleneck is originally framed as a supervised method targeting the relevant information regarding a task variable and the beta-VAE extends naturally to this setting through the same variational objective [99]. Consequently, RDT enables incorporating the effects of task demands on the learned representation in a principled way through the specification of the distortion function. While introducing such demands into the distortion would allow an exploration of further aspects of memory distortions, we left these for future work, restricting ourselves to the unsupervised version in our study. Bates and Jacobs propose another method for incorporating task-variables in their model by extending the unsupervised component with a decision network, allowing task-variables to affect the latent representation. Among other applications, they use these task variables to model category bias, however they point out that such bias can also appear due solely to categorical structure present in the data distribution. This latter explanation is what we appeal to in our study, although we agree that the latent representations in semantic memory are presumably also shaped by task objectives. We believe that further progress in machine learning in the area of learning generative models will allow lifting current limitations and will provide the background for a fully consistent model of memory.

Compression, and more specifically RDT has been proposed as a framework for an ideal observer analysis in visual working memory and perception tasks [134, 135, 136]. In Sims et al. [134] they experimentally demonstrate RDT's prediction that if memory is optimised for the statistics of stimuli learned in the course of the experiment, recall should be less accurate in case the distri-

bution of stimuli has high variance as opposed to a low variance condition. In Sims et al. [135] they infer the distortion function in a bottom-up fashion from behavioural data. One major point of contrast between this approach and ours is that instead of inferring the distortion function from behaviour, in our study the distortion function is implicitly defined by the inductive biases inherent in using latent variable generative model used for compression. A second point of contrast is that in all of these works, the authors apply RDT to low dimensional perceptual tasks with simple statistics, where optimal encodings are feasible to compute directly. In our approach it is the complex, high-dimensional and strongly structured nature of input statistics that necessitates the use of generative models which in turn define the distortion function. In [136] they analyse colour perception and memory and in addition to a low-level distortion term measured in pixel space they introduce an additional term which penalises distortions that result in the reconstruction crossing colour category boundaries. The additional term in the distortion introduces a category dependence of reconstruction similarly to the category related distortions we have modelled with the sketch-VAE, however they did not explore contextual effects in reconstruction. We argue that these perceptually simple tasks, while allowing for quantitative comparison of predictions with experimental data, are less capable of inducing the kinds of distortions that we are concerned with here, as the need for generative models is most crucial when the input distribution is complex, high-dimensional and strongly structured.

Our work is closely related to Bayesian account of reconstructive memory approach of Hemmer et al. [12] where they provide a normative method for combining episodic and gist information available in memory through an optimal Bayesian decoder. They assume that stored values for features in the memory trace are noisy versions of the observed values. These noisy values are then combined with feature priors through Bayes' rule, which reduces noise in the memory through exploiting prior knowledge. This decoding step is similar to reconstructing the observation in a generative model conditioned on inferred latents in our approach, however they do not consider the encoding step as part of the same process which should be optimised. In semantic compression, features are prioritised according to the distortion function and the optimisation also concerns what information should be kept as part of the trace. As a re-

sult, the amount of memory noise can vary as a function of how important each feature is in relation to others and the memory constraints, which is a fundamental difference from the setting considered in Hemmer et al. [12] (e.g. the sketch-VAE automatically chooses a trade-off in accuracy between features such as the presented glasses' shape, the angle of the rims and the length of the bridge connecting the rims). Hemmer & Steyvers [11] use a dual-route generative model to explain the effect of semantic memory in a scene recall task, but they do not relate their method to compression. Their topic model of semantic memory is very similar to our text-VAE for certain settings of beta, and they have a parameter corresponding to capacity. However, this capacity parameter only affects the episodic route and does not affect the representation of the semantic model. A crucial contribution of the RD perspective is that it provides a principled way of changing the representation as a function of available capacity. As a result, in our treatment dual routes are not required, since RDT provides a continuous trade-off between episodic-like and semantic-like memory traces. Furthermore, as the reconstructive ability of the semantic route is not affected by capacity in their model, we believe that explaining delayed recall results of very long delays such as in Seamon et al. [114] would require even further assumptions. A trade-off similar to that implied by RDT but without establishing a formal link to the theory of lossy compression has been formulated in the context of communicative interaction and leads to the emergence of semantic categories [137].

Human memory has a remarkable capacity to adaptively support decisions in a versatile environment but it also displays a rich array of distortions [87, 138]. These systematic errors have the potential to shed light on the design principles of our memory systems [86]. Performing complex tasks by agents suggests that various computations can be supported by episodic and semantic memory systems [139, 140, 61, 141]. Accordingly, memory distortions have also been linked to different computational processes. In particular, besides diminishing resources, other normative arguments have been made to understand various aspects of time-dependent deterioration of memories. Interference of memories from novel experiences has been linked to the flexibility of the represented model [73, 142] and regularisation was proposed as a normative principle, which could help preventing overfitting [143]. Dynamics of the environment

has been linked to adaptive forgetting rates [144, 145] and destabilisation of earlier memories after their reactivation has been linked to model update [146]. Normative models of memory distortions fall into two broad categories. One family of studies explored how the usage of latent variables to encode experiences, i.e. performing inference in the internal model, introduces systematic distortions [147, 11, 148]. Another family of studies explored how updating this internal model of the environment, i.e. learning the internal model, leads to various forms of memory distortions [149, 150, 54]. Common in all these models is that an internal model of the environment is assumed to underlie learning and inference and these internal models are described in terms of a generative model of the environment. This indicates that capitalising on more complex generative models capable of learning representations of more naturalistic data in multiple domains can contribute to a deeper understanding of memory dynamics in natural environments.

Chapter 4

Implementation level analyses

In this chapter, we will explore two case studies that extend the top-down approach of previous chapters by constructing mechanistic models on the level of neural networks. First, we examine how the idea of semantic compression and its realisation through hierarchical generative models can be related to models of the visual cortex, as well as how it can explain certain features of measurements of neural activity in V1 and V2. Second, we present an approach for implementing structural knowledge, such as the partitioned task representations discussed in Chapter 2, in neural networks and learning this mechanism in the context of the behavioural experiment analysed in Section 2.5.

4.1 Hierarchical semantic compression in the visual cortex

Sensory processing produces hierarchical representations, which according to the semantic compression hypothesis, extract increasingly behaviourally relevant quantities from raw stimuli. Predictions of neural activity in hierarchical systems are most often made in supervised deterministic models, while probabilistic generative models provide a more complete unifying view of sensory perception (Section 1.2.2). Whether unsupervised generative models trained on naturalistic stimuli give rise to representational layers of semantically interpretable quantities is yet unresolved, as is whether such representations can predict properties of neural responses in early vision. We use hierarchical variational autoencoders to learn a representation with graded compression levels

from natural images, which exhibits variance according to perceptually relevant texture categories. We predict measures of neural response statistics by assessing the posterior distribution of latent variables in response to texture stimuli. Experimental results show that linearly decodable information about stimulus identity is lost in the secondary visual cortex while information is gained about texture type, which behaviour is reproduced by the representational layers of our model. Deep generative models fitted to natural stimuli open up opportunities to investigate perceptual top-down effects, uncertainty representations along the visual hierarchy, and contributions of recognition and generative components to neural responses.

Animals need to discard the bulk of information acquired through their sensory organs to obtain the tiny portion that will be used for present or future decisions in various tasks. The semantic compression hypothesis [55] proposes that the efficient way to lose information is to encode the stimulus through the latent variables in a model of the environment. In the ventral stream of the visual cortex, compression is realised in a hierarchical manner where the sequence of compression steps culminates in the recognition of objects and concepts where variance along complex variables such as pose, lighting, and scale are discarded. The Bayesian brain hypothesis suggests that successive layers of representation correspond to latent variables of a hierarchical generative model [124]. We propose that applying the semantic compression hypothesis to a hierarchical Bayesian model results in a sequence of representational layers that extract increasingly abstract descriptors of the observation from the statistical properties of the stimulus (Fig. 4.1). Consequently, representations will be invariant to increasingly complex transformations at each layer. For example, as depicted in Fig. 4.1, when presented by an animal fur pattern, conditioning on the lowest-level inferred latents we can generate the same pattern with different observation conditions (such as lighting), on mid-level latents, different samples from the same type of fur pattern, and on higher-level latents, different types of fur patterns. Measurements of auditory perception have shown the extraction of summary statistics from complex stimuli [117] in a way compatible with semantic compression. Here we aim to show that unsupervised hierarchical models also extract semantically relevant latent variables in the visual domain.

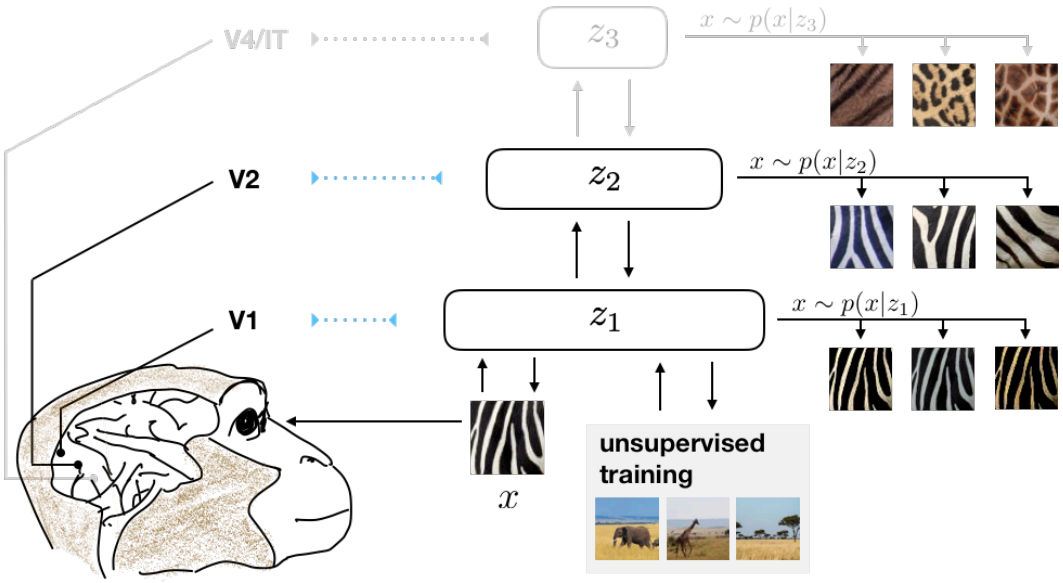


Figure 4.1: Vision as hierarchical inference. A probabilistic generative model with multiple layers of latent variables ($z_{1..3}$) is trained to compress natural images efficiently (middle). When a specific stimulus (x) is presented to the model, we can infer the corresponding latent representation in each layer. We can condition the generative model to the inferred representation on any specific layer and generate samples from the model that keep the information represented at the given layer and sample lower-level details from the learned distributions (right). Layers of representation in the model can be used to make predictions about measurements from different areas of the ventral stream of the visual cortex (left).

Assessing representations in the visual hierarchy

What quantities influence the activity of neurons in various parts of the visual hierarchy beyond the primary visual cortex (V1) is a question far from settled. Recent studies characterise mid-level sensitivities in the secondary visual cortex (V2) [151]. How such sensitivities constitute a representation can be defined in multiple ways, the simplest of which is linear decodability. Successive processing areas implementing increasing linear decodability of behaviourally relevant quantities is proposed in the hypothesis of representational untangling in higher-level visual areas [152]. Furthermore, recent evidence suggests that information related to texture categories is available linearly in V2 but not in V1, while linear information about the stimulus identity available linearly in V1 is lost at V2 [125], indicating different degrees of compression being implemented in the early visual hierarchy as well. Here we set out to investi-

gate if semantic compression in a generative model trained on natural images reproduces this signature.

Predicting neural responses in hierarchical systems

Recent studies demonstrated impressive performance on predicting neural activity in hierarchical systems [153, 154]. These models rely on feed-forward deep networks trained to classify images. However, recent evidence suggests that increasing predictive performance will require the consideration of top-down effects [155], which have also been shown to play a role in cortical computations experimentally [156]. Probabilistic generative models are well suited to describe such effects [156], and have been used to predict response statistics in early vision [157, 158, 159]. As opposed to feed-forward networks, hierarchical generative models are trained in an unsupervised way, trying to learn the distribution of inputs as well as possible given capacity constraints instead of trying to perform a specific task well, which is exactly what we expect different layers of representations to do if they are to compress inputs to different degrees. The question of whether semantic compression is a product of task-training or obtainable in an unsupervised way remains open.

There are a number of architectural choices one has to make when building a generative model. The lowest level of visual cortical representations is suggested to be close to linear by [160], formulated as a generative model by [161]. Beyond V1 we obviously need nonlinear computations, but the constraints on the kind of generative model that would capture this computation are not well characterised, thus warranting the application of a generic machine learning model implementing hierarchical inference. In this study we propose a hierarchical probabilistic generative model fitted to natural stimuli, producing multiple layers of increasingly compressed representations, suitable to make predictions about statistical properties of neural activity in visual cortical areas in response to specific images.

Texture families to probe layers of representation

In order to probe hierarchical representations, we need stimulus sets of compositional nature, such that low-level local features of the image are or-

ganised to define an abstract property for the stimulus, which can be treated as a categorical label. Recent results suggest that texture is a relevant abstraction for the secondary visual cortex [151]. Texture images can be synthesised using photographs of natural textures [162], enabling us to produce a large number of samples from the same texture family (Fig. 4.2A). Such texture families are well suited to test semantic compression through linear decodability, since the average image of a family is always zero, all family-specific information being present in higher-order pixel statistics, as opposed to e.g. the digit categories of MNIST which are decodable linearly from the pixel space (Fig. 4.2B).

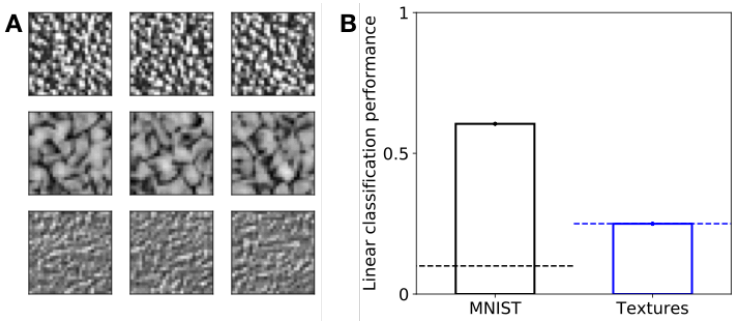


Figure 4.2: **A:** Samples from three texture families used to evaluate our models. **B:** The digit categories of MNIST are linearly decodable from the data space, while texture families are not characterised by different average pixel patterns, thus are not decodable linearly. Dashed lines indicate chance values, while bars and error bars represent mean and S.E.M. of cross-validation folds respectively.

Hierarchical variational autoencoders

Bayesian inference in hierarchical models is computationally intensive, thus the brain is expected to implement efficient approximate solutions. Variational methods use tractable distributions to infer the posterior of latent variables. Recognition-generative models, such as variational autoencoders (VAE) use an explicit feed-forward model to implement variational inference [163, 93].

VAEs have been used to describe semantic compression using a capacity parameter to balance the fidelity and the bandwidth of the latent representation [100]. They have been demonstrated to capture the abstraction of the digit category in MNIST, but not in more complex categorical stimuli [164].

A natural extension of the VAE model family is to define hierarchical layers of latent representations in order to capture the stimulus statistics at different

levels of abstraction, such as in Fig. 4.1. Learning such hierarchical representations is a nontrivial problem, for which multiple proposed solutions exist. One of these is the Ladder Variational Autoencoder (LVAE) [165], which uses a direct feed-forward mapping from the stimulus to all latent layers during inference, allowing for the efficient learning of latent hierarchies while introducing no additional computational steps into the generative model. We used LVAEs as models of the representational hierarchy in the ventral stream, fitting them to naturalistic stimulus statistics and then presenting them texture stimuli to compare properties of the inferred representations to those measured from the visual cortex.

Results

We fitted a two-layer LVAE to whitened natural image patches of 16x16 pixels obtained from the van Hateren dataset [166] (shown in Fig. 4.3E). The architecture consisted of 20 and 5 stochastic units in the two latent layers. The lower level representation was connected to the stimulus through a linear encoder and decoder. The second layer was connected to the first using two densely connected ReLU layers of 32 units each and a batch normalisation layer both in the encoder and the decoder. The observation noise was fixed at 0.1. The parameters of the model were fitted to the natural patches using the Adam optimiser for 60 epochs with a learning rate of 0.001, using a burn-in period [165] of 10 epochs.

As we wanted to compare the properties of the learned representations to those measured in macaques by [125], we inferred the latent representation of texture stimuli shown in Fig. 4.2A. We constructed linear mixture of Gaussian decoders both to distinguish between the latent representations of specific stimuli and the families they were sampled from, using the representations from both latent layers of the LVAE (Fig. 4.3B). The performance of the decoders was calculated as the cross-validated hit rate for either 4 samples from the same family or 4 from different families. The performances were plotted against each other to contrast the properties of the representations learned in the two layers (Fig. 4.3F). The first layer could be used much better to recover the identity of the stimulus. Most of this information was lost at the

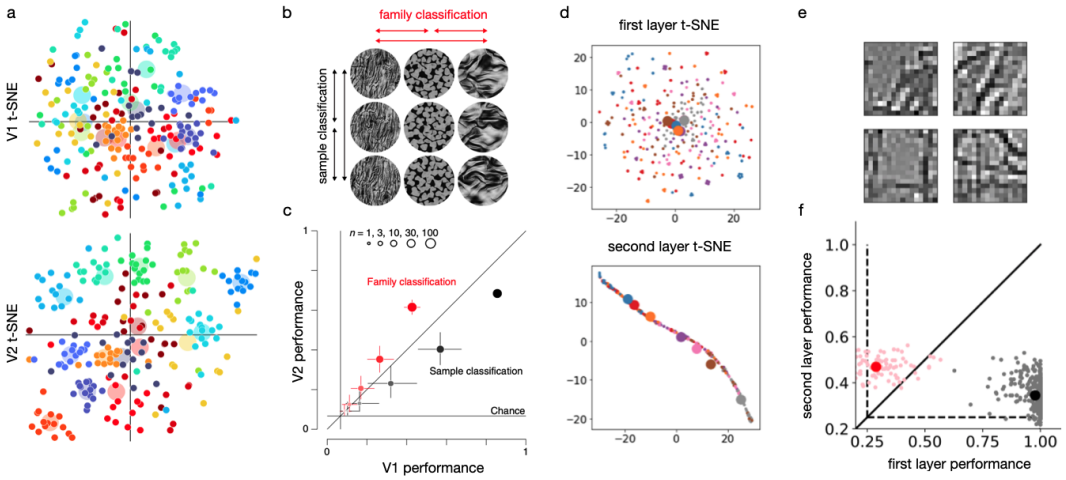


Figure 4.3: Comparison of neural representations and those learned by a two-layer LVAE from natural images. **A,** V1 (top) and V2 (bottom) population response to each visual texture stimulus, displayed in a 2D coordinate system that captures the responses of 102 V1 and 103 V2 neurons, computed using t-SNE. Each point represents one texture image, with colour indicating the texture family. Panels A-C adapted from Ziemba et al. [125]. **B,** Schematic of sample (black) and family (red) classification. **C,** Comparison of proportion of correct classification of V1 and V2 populations for family classification (red) and sample classification (black). **D,** Same as panel A but for the two layers of the LVAE model, large dots indicate the mean of each family. Samples from the same family are more clustered together in the second layer. **E,** Whitened natural patches used for training the model. **F,** The decodability of the stimulus identity of texture stimuli of the kind shown in Fig. 4.2 (grey dots) and the family they are sampled from (pink dots). Red and black dots represent the mean of all the decoding comparisons and dashed lines represent chance levels. The representation learned in the first layer of the LVAE contains more information about the identity of the stimulus, while the second layer contains more about the family (Cf. panel C).

second layer, while making the family more linearly decodable. This result is in accordance with the findings from macaque V1 and V2 (Fig. 4.3C).

Semantic compression can be probed using nonlinear read-out instead of a linear one as well. We used the t-SNE nonlinear embedding method to show that the second-layer representation of texture images is more clustered according to family membership (Fig. 4.3D), similarly to measurements from macaque V1 and V2 (Fig. 4.3A).

We explored which architectural choices are essential to produce the results we demonstrate. An indispensable feature of the model is the increasingly compressed representation in the layers. However, the compression levels can

be achieved by controlling the information capacity of the layers in multiple ways, such as the dimensionality of the layers, but also the expressive power of the encoders and decoders used in them. The latter property opens up an avenue to train models of much higher latent dimensions with similar semantic compression properties as ours.

Visualising semantic compression

The learned representations that reproduce experimentally measured untangling effects are expected to compress the stimuli at different semantic levels, as in Fig 4.1. Since we learn a model of natural images, a high number of latent units would be necessary to learn all the factors of variance that include the ones directly relevant to texture samples, making the levels of variation easily observable visually. Instead of training such a model, we retrain an LVAE using the texture stimuli, directly producing the subset of latents that describe these stimuli in particular. We then infer the latent representation in each layer in response to specific textures, and condition the generative model on the inferred representation in each layer. We indeed observe that conditioning on the lower layer produces samples that reproduce the particular content of the input image and differ only due to the observation noise which is independent across pixels. Conversely, conditioning the generative model on the higher layer produces samples that come from the same texture family as the input, but vary in terms of the particular realisation of the texture (Fig. 4.4). Variational autoencoders implement hierarchical Bayesian inference producing a series of increasingly compressed representation of the input. When trained on natural images in an unsupervised way, they reproduce representational untangling of texture stimuli similarly to the visual cortex of macaques, while the learned generative model exhibits variations in the successive representational layers corresponding to perceptual categories.

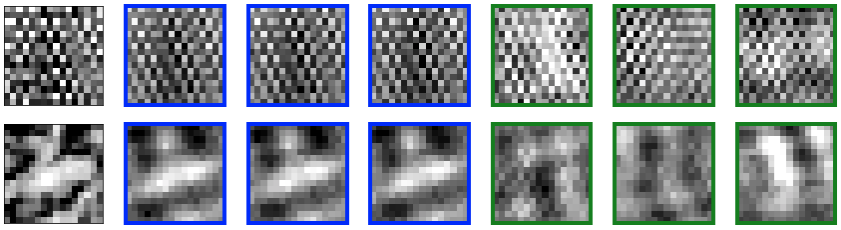


Figure 4.4: Layer-conditional reconstructions from a two-layer LVAE trained on texture stimuli. For two example stimuli (no border) we take three samples from the posterior distribution of all latent variables, and generate synthetic stimuli conditioning on the latent samples in the first (blue border) and the second (green border) layer. First-layer generated samples differ only in terms of pixel noise, while second-layer samples are different instances of the same texture family.

4.2 Order effects and knowledge partitioning in neural networks

As we noted in Section 2.5, it is a well-known problem in machine learning that artificial neural networks (ANNs) experience catastrophic forgetting under continual learning. This means that when these networks are trained sequentially on multiple tasks, they can only retain performance on the most recent task [73]. However, the same networks can learn multiple tasks if the training data is presented in a randomly interleaved manner. Interestingly, the opposite has been observed in humans, who find it easier to identify the correct task structure when training data is presented in a sequential or blocked format [74]. This sensitivity to order in standard ANNs also stands in contrast to our model of online structure learning presented in Chapter 2, which exhibits a pattern similar to the human data. Here, we present a proposed mechanism for how task structure, and specifically the knowledge partitioning by task that we have relied on in the algorithmic-level model in Chapter 2 can be achieved in ANNs and learned in the tree planting experiment of Flesch et al. [74]. Additionally, we demonstrate how an expectation of a slowly changing environment can result in confusion between the two contexts and impair the discovery of the correct model structure in the interleaved setting.

Methods

In this work, we used the same context-dependent decision task from Flesch et al. [74] that we described in detail in Section 2.5. In this task, stimuli varied continuously along two independent feature dimensions, and only one of the two dimensions was relevant in each context. While the original experiment used high-dimensional fractal tree images that varied in their density of leaves and branches (as shown in Fig. 2.7), we simplified the neural network version by replacing them with stylised images of Gaussian blobs whose x and y locations varied (Fig 4.5A). We have also verified that the analysis applies to ANNs trained on subsampled versions of the tree images (not shown; for details, see the supplementary information of the full paper [59]).

For all of the analyses presented, we used three-layer perceptrons with ReLU nonlinearities as our model of a standard ANN, also referred to as a vanilla NN. The output of the network was a single neuron corresponding to the expected reward, given that the tree observed in the trial would be planted. The inputs to the network were the stylised images, concatenated with the context encoded as a one-hot vector, which represented the background for the tree image in the original version.

Context-modulated gating of task representations

When learning the tree task with interleaved contexts over trials, the standard ANN learns a partitioning of the hidden layer (Fig. 4.5B&C) via orthogonal patterns of weights from context input nodes to hidden units. This allows learning of the relevant mappings for the two tasks to occur in orthogonal subspaces (not shown, for details see [59]). At a mechanistic level, the network learns anti-correlated context weights, so that the active context activates hidden units which carry information regarding relevant feature dimensions and inhibits units carrying task irrelevant information. This inhibition maps the irrelevant units to negative values, which are subsequently filtered out by the ReLU nonlinearity (Fig. 4.5B, left). Importantly, because units that are active only in the first context are gated off during the second task, the gradient signal does not backpropagate and disrupt the representation learned for the first task (4.5b, right). In this coding scheme, the context units act as gating

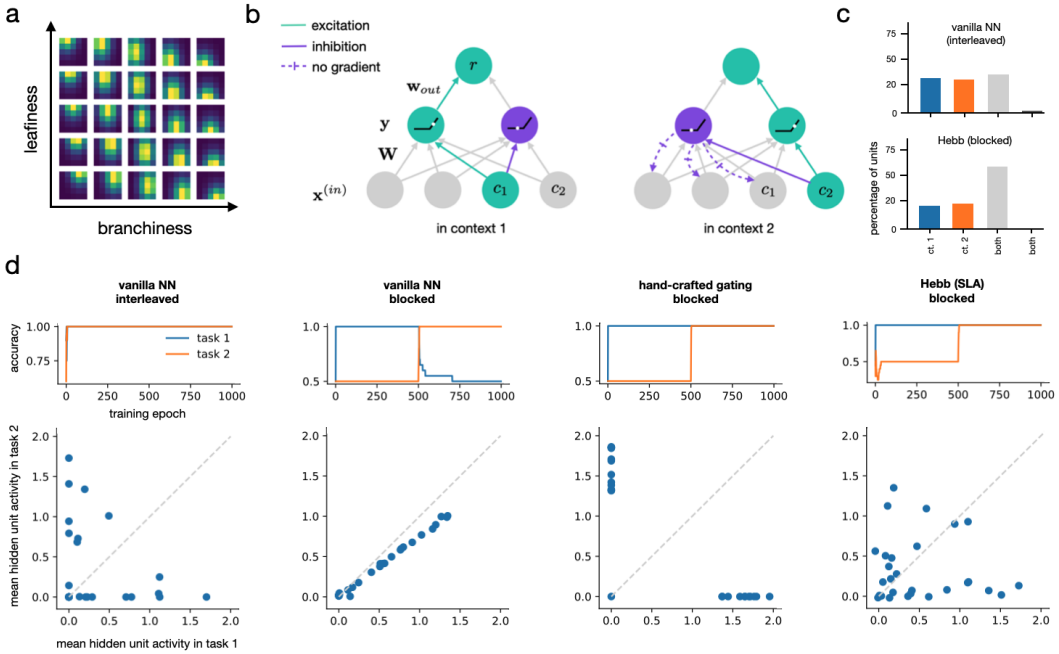


Figure 4.5: Interleaved vs blocked training in vanilla, gated and hebbian networks. **a,** Simplified Gaussian blob input for each value of the latent features. **b,** Schematic illustration of gating mechanism. The active context unit (green $x^{(in)}$) increases the activation of the context relevant hidden unit (green y) and pushes it into the linear interval of the ReLU activation function, letting this activation pass on to the output. Conversely, the same unit inhibits the irrelevant hidden unit (purple y). In the second context (right), the other context unit is active, and the output is defined by a different hidden unit. The ReLU protects the weights leading into the inactive hidden units in a given context. **C:** Context selectivity of hidden units in the standard version of the NN (top) and the NN augmented with the Hebbian update rule (bottom). **D:** Accuracy of each model over training epochs (top) and trial-averaged hidden unit activity in each context. Points lying on the diagonal correspond to neurons that are active to the same extent (on average) in both contexts.

variables that selectively allow the propagation of task relevant information. To demonstrate that this coding scheme is sufficient for multitask learning, we set the context weights by hand and show that it allows the network to avoid catastrophic forgetting even under the blocked schedule (Fig. 4.5D).

While the hand-crafted context weights prevent catastrophic forgetting in the interleaved setting, the question of how such a gating scheme can be learned in the blocked case remains unanswered. We hypothesised that an associative (or Hebbian) learning rule, in which simultaneous activation of cells results in strengthened synaptic weights, would lead to an association between task relevant hidden units and the currently active context. Conversely, hidden units

carrying task-irrelevant information would be anti-correlated to the active context and thus their weights would be weakened. To test this hypothesis, we employed a variant of Hebbian learning called the Subspace Learning Algorithm (SLA) [167]. On each trial, we apply both a standard SGD update to learn the task and a Hebbian update to learn the context weighting:

$$\Delta \mathbf{w}_i = \eta x_{i(t)}^{(h)} [\mathbf{x}_t^{(in)} - \mathbf{W}\mathbf{x}^{(h)}],$$

where $\mathbf{x}^{(h)} = \mathbf{W}^T \mathbf{x}^{(in)}$ is the input to the hidden layer before the ReLU nonlinearity is applied. Similarly to the Generalised Hebbian Algorithm (GHA), SLA learns the principal subspace of the input data. If the context signal is sufficiently large, one of the first eigenvectors will correspond to the context variance in the input. However in the case of GHA, since it learns the principal components, this variance will be represented in a single basis vector, consequently affecting only a single hidden neuron. The SLA learns the same subspace but on a randomly rotated basis, which distributes the context variance and leads to the desired anti-correlated weight structure. We have found that similarly to the manually gated network, the SLA network assigned a subset of hidden units to each context and therefore did not suffer from catastrophic forgetting under blocked training (Fig. 4.5D).

Expectation of correlated input impairs interleaved performance

While context-modulated gating protects against catastrophic forgetting and results in maintained task performance under both interleaved and blocked training, humans find interleaved training harder. One possibility that we introduced in section 2.5 is that the interleaved setting might predispose subjects to a simpler model structure where the decision boundary is shared between the two gardens. We propose that a second possibility is that humans may have an expectation, based on natural statistics, that task identity changes slowly over time. Since the tree background can be directly observed by the subjects, this implies that they assume that the true context is connected to the background in a stochastic manner. In this case, the true context is a latent state variable that needs to be inferred from the observation. On a computational level, this task can be formalised as inferring the true task identity in a Hidden Markov

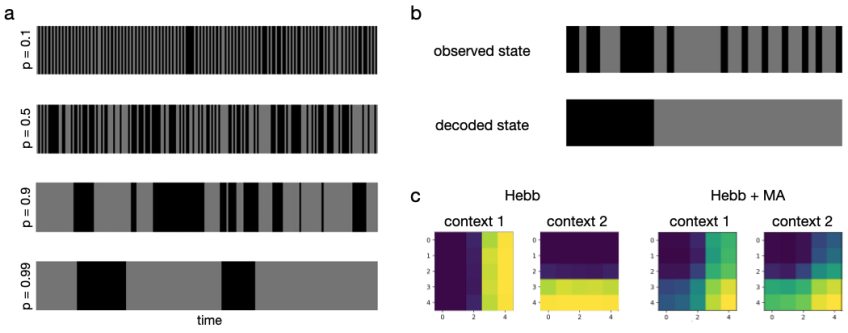


Figure 4.6: Expectation of correlated input and sluggish network results. **A,** Synthetic data generated from a HMM, where the two colours correspond to the two possible contexts. With increasing probability of self-transitions p, the latent states are increasingly auto-correlated. **B,** Using a high value of p, the decoded state is smoothed relative to the observed state. **C,** Choice probability plots under interleaved training for the Hebbian (SLA) network and the Hebbian (SLA) network with time-averaging over the context unit.

Model (HMM), where a prior expectation of slowly changing context translates to a high prior probability for self-transitions (Fig. 4.6A). This 'sticky' prior uses information from previous trials and smooths the observed transitions in the observations (Fig. 4.6B).

In our ANN model, we have incorporated the idea of slowness or stickiness through a temporal smoothing of the context units' activity across trials, as suggested by Földiák, 1991 [168]. This smoothing led to reduced performance on interleaved data and the observed intrusion of irrelevant dimensions, similar to what has been observed in humans (Fig. 4.6B). Note that the model fits use a variant of the model where the activity of the context unit was an explonential moving average (EMA) of the context input, and the SLA rule was replaced with a simpler variant called Oja's rule, applied solely between the context and the hidden units. To explain the relative benefits of blocked over interleaved learning seen in humans, we re-analysed data from Flesch et al., 2018 [74]. We quantified how well networks with and without gating and smoothing procedures could explain human choice patterns. We have found that the network with the addition of the Hebbian mechanism and slow context prior provided a better explanation of the data than its vanilla counterpart (Fig. 4.7).

Our mechanistic model explains the observed advantage of blocked over interleaved learning in humans, using insights from neural network research

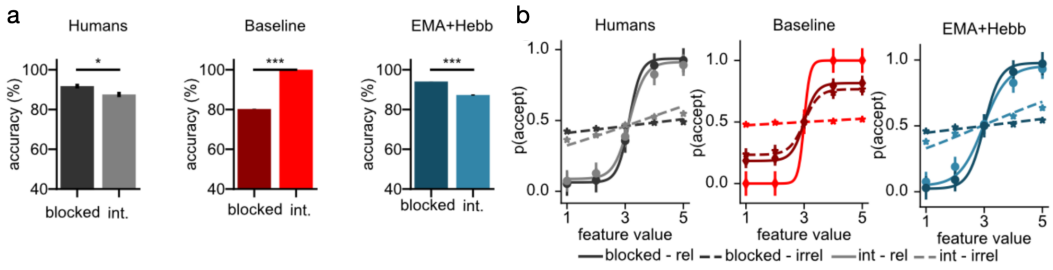


Figure 4.7: Modelling the benefit of blocked over interleaved training observed in data from human participants. **A,** Test phase accuracy. Humans perform better under blocked compared to interleaved training, while the baseline model performs better under interleaved training, due to catastrophic forgetting. Our Hebbian network with time-averaged context units performs better under blocked training. **B,** Sigmoidal fits of choices made by human participants, the baseline network, and our network. Our network recreates the intrusion of irrelevant dimensions observed under interleaved, but less so under blocked training. Analysis performed by Timo Flesch, figure reprinted from Flesch & Nagy et al., 2023. [59]

to demonstrate a biologically-plausible approach for acquiring orthogonal task representations. Furthermore, we have proposed a key difference in the assumptions underlying the learning strategies of humans as opposed to standard ANN learning algorithms and validated our approach by fitting the network responses to human choice data. Our findings provide novel insights into the learning mechanisms of artificial neural networks and puts forward a specific mechanism for continual learning in humans.

Chapter 5

Summary and conclusions

In this thesis, we aimed to provide a computational level account of long-term memory in humans and other intelligent agents operating under similar resource constraints. Our analysis was based on the premise that a primary role of memory is in enabling the brain to predict future sensory inputs. A key piece of this challenge is to extract the information needed to construct and update an internal model of the environment, which has been identified as the main goal of semantic memory.

In the first study, we aimed to demonstrate why an episodic memory system is necessary for a learning agent and how it complements the semantic memory system. We argued that online structure learning in open-ended model spaces results in either tracking a possibly infinite number of competing models or retaining the entire observation history, none of which is feasible for a bounded-resource memory system. On two simple model systems we demonstrated that i) a semantic-only learning agent under resource constraints is biased towards its current model estimate as evidence for alternative model structures is lost upon incorporating observations into the posterior, ii) the addition of a small episodic buffer to the semantic-only learning agent can effectively circumvent this challenge of online model selection by allowing statistical power of observations to accumulate, iii) the beneficial effect of the episodic buffer increases if episodes that have large information content regarding the posterior are preferentially retained, both learners are sensitive to the order in which observations are presented, exhibiting a qualitatively similar pattern to human subjects in the experiment of Flesch et al.[74].

In the second study, we examined the opposite perspective, investigating how semantic memory can aid the functioning of episodic memory. We have demonstrated that casting the compression of episodic memories in the normative framework of rate distortion theory and exploiting the probabilistic generative model of the environment maintained in semantic memory for generative compression of episodes results in systematic compression artefacts that are also characteristic of human memory distortions. Semantic compression therefore offers a unifying explanation for a wide range of memory distortions observed in the experimental literature. Specifically, we demonstrated on naturalistic domains of chess positions, human drawings and natural text that semantic compression predicts i) increasing recall accuracy with increasing domain knowledge of the subject and increasing congruence of observations with natural statistics ii) gist-based distortions in recall and recognition where low-level details of the observation are resampled in a way that is consistent with high-level latent variables and that iii) progressive loss of episodic detail from memory traces as the rate distortion trade-off controlled by the Lagrange multiplier β is varied, consistent with forgetting in humans as the interval between presentation and recall increases.

Finally, we have extended our computational and algorithmic level analyses to the question of how these algorithms can be implemented in the brain and artificial neural networks. We implemented semantic compression and specifically movement along the rate-distortion curve using a hierarchical generative model and demonstrated that at higher levels in the hierarchy, where the observation is compressed at a lower rate, texture-family relevant information is maintained while sample-specific information is lost. We used these results to model measurements from the macaque visual cortex where it was found by Ziemba et al. [125] that sample identity is easier to decode than texture family from V1, whereas this relation reverses in V2. In a separate artificial neural network model, we have demonstrated that task representations can be partitioned in a context-modulated gating scheme, which makes the network robust to blocked training. We have shown that this gating scheme can be learned in the same context-switching experiment as in the first study if context neurons are associated with hidden neurons carrying context relevant information by the addition of a Hebbian learning rule called the Subspace

Learning Algorithm. We further demonstrated that an expectation of slowly changing or auto-correlated context, implemented as time averaged context unit activity in the network, leads to the intrusion of context-irrelevant information in the interleaved setting and can explain the order sensitivity observed in human data.

Appendix A

Supplementary information

A.1 Episodic memory

A.1.1 Inference and learning in mixture of Gaussians

Notation

T is the number of observations, K is the number of components in the mixture, z is a one-hot binary vector, γ_k are the prior assignment probabilities.

Due to the graphical model of the Gaussian mixture, the likelihood factorises according to:

$$p(\mathcal{D}, \mu, \mathbf{z}) = \prod_{i=1}^{T} p(x_i|\mu, z_i)\ p(\mu|z)p(z)$$

Inference of assignments with unknown parameters

After observing the first data point x_0,

$$p(x_0, \mu, z) = \prod_{k}^{K} \mathcal{N}(x_0|\mu_k, \sigma_k)^{z_k} \cdot \mathcal{N}(\mu_k|\mu_0, \sigma_0) \cdot \gamma_k$$

Where we can get the posterior over assignments by using the definition of conditional probability and marginalising over μ:

$$p(z|x) = \frac{p(x, z)}{\sum_z p(x, z)} = \frac{\int p(x, \mu, z) d\mu}{\sum_z \int p(x, \mu, z) d\mu}$$

If we then substitute in the full joint distribution we get

$$p(z|x) = \frac{\int \prod_k^K \mathcal{N}(x_i|\mu_k, \sigma_k)^{z_k} \cdot \mathcal{N}(\mu_k|\mu_0, \sigma_0) \cdot \gamma_k d\mu}{\sum_z \int \prod_k^K \mathcal{N}(x_i|\mu_k, \sigma_k)^{z_k} \cdot \mathcal{N}(\mu_k|\mu_0, \sigma_0) \cdot \gamma_k d\mu} = \frac{\mathbb{N}}{\mathbb{D}}$$

Looking first at the numerator

$$\mathbb{N} = \gamma_k \int d\mu_1 ... \int d\mu_n \mathcal{N}(x_i|\mu_1, \sigma_1)^{z_1} \mathcal{N}(\mu_1|\mu_0, \sigma_0) ... \mathcal{N}(x_i|\mu_K, \sigma_K)^{z_K} \mathcal{N}(\mu_K|\mu_0, \sigma_0),$$

we have to multiply Gaussians in each dimension, which result in the same parametric form,

$$\mathcal{N}(x|\mu, \sigma)\mathcal{N}(\mu|u_0, \sigma_0) = c \cdot \mathcal{N}(\mu|\mu', \sigma'),$$

but with new parameters

$$c = \mathcal{N}(x|\mu_0, \sqrt{\sigma^2 + \sigma_0^2});\ \mu' = \frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2};\ \sigma' = \frac{\sigma\sigma_0}{\sqrt{\sigma^2 + \sigma_0^2}}.$$

We may then notice that we only have to take integrals of Gaussian pdf-s multiplied by the normalising constants c_k, therefore the result can be written simply as

$$\mathbb{N} = \gamma^K \begin{cases} c_1 & z_1 = 1 \\ 1 & z_1 = 0 \end{cases} ... \begin{cases} c_K & z_K = 1 \\ 1 & z_K = 0 \end{cases} = \gamma^K \prod_k^K c_k^{z_k},$$

and consequently the full fraction is:

$$p(z|x) = \frac{\prod_k^K c_k^{z_k}}{\sum_z \prod_k^K c_k^{z_k}}$$

Parameter learning

The posterior over the Gaussian means is

$$P_m(\mu|x) = \frac{P_m(x|\mu)P_m(\mu)}{\int P_m(x|\mu)P_m(\mu)d\mu},$$

where for $k = 1$ and Gaussian priors we again only have to multiply Gaussian pdf-s and the posterior is the same form but with updated parameters (con-

jugate prior). If $k > 1$, then we may note that the result takes the form of a mixture of such conjugate updates

$$P_m(\mu|x) = \frac{\overbrace{\sum_{k=1}^{K} \frac{1}{K}\mathcal{N}(x|\mu_k, \sigma)}^{P_m(x|\mu)} \prod_{l=1}^{K} \mathcal{N}(\mu_l|\mu_0, \sigma_0)}{\int \sum_{k=1}^{K} \frac{1}{K}\mathcal{N}(x|\mu_k, \sigma) \prod_{l=1}^{K} \mathcal{N}(\mu_l|\mu_0, \sigma_0) d\mu}.$$

As the number of observations increases, the number of terms in the posterior grow exponentially as K^T, however each term can be computed analytically. Therefore we can apply a Rao-Blackwellised particle filter, where we maintain a fixed number of likely assignments (particles) but for each assignment we can perform analytical updates along the remaining dimensions of the posterior.

A.2 Semantic compression

A.2.1 Chess-VAE

We have trained a beta-VAE on games downloaded from the FICS (Free Internet Chess Server) Games Database, containing hundreds of thousands of recorded online games. Chess configurations were represented as a one-hot vector for each one of 64 squares on the chessboard, with each 13 dimensional one-hot vector specifying the chess piece or the lack of a chess piece in that square. The decoder output was a categorical distribution for each square, which represented the probabilities of possible chess pieces on the particular position. Both the encoder and decoder of the chess-VAE consisted of two dense layers with 3000 units and sigmoid nonlinearity, followed by a linear transformation. The prior over the continuous 64 dimensional latent space was a Normal distribution with an identity covariance matrix. Reconstruction was modelled by conditioning the model on a chess configuration, inferring the latent representation, then taking the MAP reconstruction of the decoder.

Since the goal was to demonstrate robust qualitative effects resulting from the theoretical framework, hyperparameters of the model were chosen so that reconstructions would fall in a regime comparable to the experimental data as opposed to fine tuning them for an accurate match. Beta is a central free

hyperparameter for the model, which was chosen to be 0.0001 so that the 'expert' model could reproduce the state of around 90% of squares correctly in the MAP estimate. Decreasing beta further did not result in flawless recall, presumably due to capacity limitations in the encoding and decoding transformations. Patterns in the presented results were relatively robust to changes in parameter β but increasing it by orders of magnitude causes both the overall accuracy and difference between the random and game conditions to decrease. Varying the size of latent space had similar effect to varying β as it modulates the capacity of the latent space. Decreasing the hidden layer widths from 3000 led to qualitatively similar results but with performance in the game board condition saturating at a lower level of around 15 successfully reconstructed pieces in the case of 1000 units and 10 in the case of 500 units. Doubling the number of training steps did not noticeably change the resulting reconstruction accuracies. Separate training and test sets were formed from all the board positions from 3000 games of the FICS Games Database. Different levels of skill were modelled by using different sized training sets: game positions were subsampled to 0.1%, 1%, 10%, and 90% of the full dataset. Training set sizes corresponding to various skill levels were also not precisely calibrated to experimental data but set to what we determined to be sensible values so that amateur players would only observe a few games and expert players would see a sufficient amount to reconstruct with comparable accuracy to the experiments. Remaining set sizes were chosen to span orders of magnitude between these two extremes. For training we have used Adam with a learning rate of 10^{-4} and a batch size of 65.

The test configurations were constructed as in Chase et al. [169]. 'Game' configurations were taken after either the 21st or the 41st move in games from a separate test set. 'Random' configurations were produced by shuffling the pieces across occupied board positions of a game setting. The number of pieces on the board were not fixed in the dataset and could vary across trials. We have followed the accuracy evaluation method proposed in the Gobet and Simon paper, where the number of correct pieces on a reconstructed board is counted.

A.2.2 Text-VAE

The beta-VAE constructed to learn the statistics of natural text was similar to the chess-VAE: we used an encoder and a decoder with two dense layers with 2000 hidden units per layer followed by sigmoid nonlinearities. Activations, $\mathbf{z}$, in the last hidden layer of the decoder are used to generate words, $\mathbf{X}$ independently through a linear transformation and a softmax nonlinearity according to

$$e_i = exp(-z^T R x_i + b_{x_i})$$

$$p_\theta(x_i|z) = \frac{e_i}{\sum_j^{|V|} e_j}$$

$$p(X|z) = \prod_i^N p_\theta(x_i|z),$$

where R is a $|z| \times |V|$ matrix that acts as a semantic word embedding, organising words into a low dimensional continuous vector space such that similarity in this space reflects semantic similarity between the words. The model had a continuous latent space with a diagonal Gaussian prior in 100 dimensions. This architecture (with minor differences) has appeared in the machine learning literature previously as the Neural Variational Document Model (NVDM) [108]. The factorisation assumption, which formulates the generation of multiple-word text as an independent process, greatly simplifies training but the consequence is that the same word can be generated multiple times in a synthesised text. Since this is a purely computational simplification that is clearly detrimental to performance in the world list recall task, we constrained reconstructions such that distinct words had to be generated. For training on both Wikipedia and the synthetic dataset described below, we have used Adam with a learning rate of 10^{-5} and a batch size of 100. We have chosen these hyperparameters based on the Miao et al. [108] paper, adjusting them for the fact that they have trained their model on significantly smaller datasets (vocabulary sizes of around 2000 and 5000 as opposed to 50000). The large dataset was required so that we have a sufficiently good approximation of language statistics, which we assessed through testing whether the associations that the original

DRM lists rely on are reliably represented in the model. Consequently we have increased the training set to the extent that was computationally feasible for us to train. Beyond these considerations no further adjustment of the model parameters was required and results are presented without fine tuning of the parameters. Due to the computational cost of training the model, we have not explored perturbations of hyperparameters extensively. We have calibrated β values such that the reconstruction probabilities for studied and critical lure words were approximately the same, which is what was observed in the original DRM experiment.

To control for the mismatch between the statistics of the training corpus and the natural vocabulary humans experience over a lifetime, we only used the word lists from the original DRM article which fulfilled the criterion that the lure word had at least 2000 occurrences in the training set. Words that appeared less than 100 times in the training set were deleted from the lists, and if these manipulations resulted in a list that was shorter than 12 words then the entire list was removed. According to these criteria we included 10 of the original word lists in the analysis (*high, rough, mountain, music, black, man, foot, king, river, soft*). As an illustration of the similarity of associations between the model and human data, 15 closest associates of the word '*music*' in the model are '*musical, album, arts, songs, sounds, pop, art, instrument, sound, musicians, progressive, disambiguation, label, composers, string*' whereas the corresponding DRM list is: '*jazz, horn, concert, orchestra, rhythm, sing, piano, band, note, instrument, art, sound, symphony, radio, melody*'.

In order to mitigate variability in the averages due to i) the low number of word lists and ii) mismatch between the statistics of the Wikipedia training set and natural text, we also built a synthetic dataset so that the performance of the text-VAE can be explored under well controlled statistics. We generated synthetic text using an artificial vocabulary and a Latent Dirichlet Allocation (LDA) topic model. The LDA generative model used a synthetic vocabulary of 1000 words with a word concentration parameter 0.1, and 10 topics with concentration parameter 0.1. These parameters were chosen such that a t-SNE embedding would largely be able to separate the main topics in each document but not perfectly. We sampled 20000 documents from the LDA models to use as a training set. In the original memory experiment, word lists were

generated by asking subjects to list their first associates to the presented lure word. Analogously, for the artificial vocabulary we trained a separate model on synthetic data generated from the LDA model and computed the 15 most similar associates based on the learned embedding matrix R. We generated such DRM-like word lists for each word in the vocabulary that occurred at least 500 times in the training set, but discarded lists that became shorter than 12 after removing infrequent (less than 500 occurrences) words. Note that this method can also be used to construct new DRM word lists based on the model trained on Wikipedia.

A.2.3 Sketch-VAE

To obtain a generative model for hand-drawn sketches we used the sketch-RNN VAE architecture [109] that was developed specifically as a generative model for sketches and is able to capture sequential dependencies in the data. Sketch-RNN represents sketches as sequences of pen strokes, which are encoded into the latent representation through a bidirectional RNN. The output of this network then parametrises a Normal distribution over the latent space. The decoder consists of an RNN conditioned on the latent vector and preceding strokes, outputting the parameters of a mixture of Gaussians which generates the next pen movement.

In order to be able to relate our analysis to the distortions observed in the Carmichael experiment, we have used hand drawn sketches from the QuickDraw dataset Ha & Eck [109], consisting of a rich set of labelled drawings depicting 345 common object categories. The experiment used ambiguous drawings that could plausibly belong to multiple categories; hence we selected category-pairs that contained a substantial number of visually similar exemplars. Although the QuickDraw dataset contains rich naturalistic samples from every category, characteristics of recording the doodles preclude a large number of object pairs from the analysis. The QuickDraw data was recorded as part of a web browser game, where subjects had 20 seconds to draw an exemplar of a given category. However, if the drawing gets to a stage where an object classifier is able to recognise it as belonging to the provided category, a new trial is initiated. As a result, the data set contains a large number of

unfinished drawings. Another limitation of the data set is that participants tend to draw prototypical exemplars of the category thus limiting the variance of the samples compared to all possible ways of drawing the object that would still be easily recognisable by human observers. This means that some of the designs appearing in the Carmichael experiment are not present in the data set and thus the model is oblivious to their interpretation. In total we have selected five object pairs (eyeglass-dumbbell, chair-bed, wheel-fan, moon-banana, pizza-wheel) and the model was trained separately for each object category, on a training set of 75000 samples per category. The reconstructions were based on samples from a separate test set. We have modelled the effect of presenting the label by conditioning the category specific generative model on the ambiguous image and using it to generate a reconstruction.

Most parameters of the sketch-rnn model are determined by the data statistics and only a small subset of the parameters is available for fine tuning. These parameters have been explored in the original publication of the paper [109]. We have only explored the rate distortion parameter β. The value of β for different time delays were chosen so that the variance in reconstructed images was roughly comparable to that observed for humans in the experimental literature, as judged by the authors.

In the quantitative analysis of feature rescaling, following Hanawalt et al. [110], we have measured the proportion of the length of the drawing and the length of the connecting line for conditional reconstructions for 50 randomly selected samples of the test set. The length of the drawing was measured between the widest extent, excluding any stems in the case of glasses. The connecting line was defined as the distance between the two circular features at the points of intersection with the connecting line. We have only included samples where both the circular features and the connecting line was recognisable for all reconstructions.

References

[1] John R. Anderson. 'Is human cognition adaptive?' In: *Behavioral and Brain Sciences* 14.3 (1991), pp. 471–485. DOI: 10.1017/S0140525X00070801.

[2] D. Marr and T. Poggio. 'From Understanding Computation to Understanding Neural Circuitry'. In: (May 1976).

[3] James L. Mcclelland, David E. Rumelhart, and PDP Research Group. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*. en. MIT Press, July 1987.

[4] Valentino Braitenberg. *Vehicles: Experiments in Synthetic Psychology*. en. MIT Press, Feb. 1986.

[5] Jeff Hawkins and Sandra Blakeslee. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*. en. Macmillan, Apr. 2007.

[6] William Bialek and Naftali Tishby. 'Predictive Information'. In: *Arxiv* (1999).

[7] Beren Millidge. 'Towards a Mathematical Theory of Abstraction'. In: *arXiv:2106.01826 [cs, stat]* (June 2021).

[8] Tomer D. Ullman et al. 'Mind Games: Game Engines as an Architecture for Intuitive Physics'. In: *Trends in Cognitive Sciences* 21.9 (2017), pp. 649–665. DOI: 10.1016/j.tics.2017.05.012.

[9] Alexander A. Alemi. 'Variational Predictive Information Bottleneck'. In: *arXiv preprint* (2019), pp. 1–6.

[10] Szabolcs Káli and Peter Dayan. 'Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions'. In: *Nature Neuroscience* 7.3 (2004), pp. 286–294. DOI: 10.1038/nn1202.

[11] Pernille Hemmer and Mark Steyvers. 'Integrating episodic and semantic information in memory for natural scenes'. In: *Proceedings 31st Annual Meeting of the Cognitive Science Society* (2009), pp. 1557–1562.

[12] Pernille Hemmer and Mark Steyvers. 'A Bayesian Account of Reconstructive Memory'. In: *Topics in Cognitive Science* 1.1 (2009), pp. 189–202. DOI: 10.1111/j.1756-8765.2008.01010.x.

[13] Gualtiero Piccinini and Corey Maley. 'Computation in Physical Systems'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021.

[14] László E. Szabó. 'Mathematical Facts in a Physicalist Ontology'. In: *Parallel Processing Letters* 22.03 (Sept. 2012), p. 1240009. DOI: 10.1142/S0129626412400099.

[15] Warren S. McCulloch and Walter Pitts. 'A logical calculus of the ideas immanent in nervous activity'. en. In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. DOI: 10.1007/BF02478259.

[16] Gualtiero Piccinini. 'The First Computational Theory of Cognition: McCulloch and Pitts's "A Logical Calculus of the Ideas Immanent in Nervous Activity"'. In: *Neurocognitive Mechanisms: Explaining Biological Cognition*. Ed. by Gualtiero Piccinini. Oxford University Press, Nov. 2020, p. 0. DOI: 10.1093/oso/9780198866282.003.0006.

[17] H. T. Siegelmann and E. D. Sontag. 'On the Computational Power of Neural Nets'. en. In: *Journal of Computer and System Sciences* 50.1 (Feb. 1995), pp. 132–150. DOI: 10.1006/jcss.1995.1013.

[18] George Boole. *The Laws of Thought (1854)*. en. Open court publishing Company, 1854.

[19] E. T. Jaynes. *Probability Theory: The Logic of Science*. en. Cambridge University Press, 1979.

[20] R. T. Cox. 'Probability, Frequency and Reasonable Expectation'. In: *American Journal of Physics* 14.1 (Jan. 1946), pp. 1–13. DOI: 10.1119/1.1990764.

[21] R. Sherman Lehman. 'On confirmation and rational betting'. en. In: *The Journal of Symbolic Logic* 20.3 (Sept. 1955), pp. 251–262. DOI: 10.2307/2268221.

[22] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. en. MIT Press, July 2009.

[23] Noah D. Goodman et al. 'Church: a language for generative models'. In: (June 2012), pp. 220–229.

[24] Tomer D Ullman and Joshua B Tenenbaum. 'Bayesian Models of Conceptual Development: Learning as Building Models of the World'. en. In: (2020).

[25] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. 'Simulation as an engine of physical scene understanding'. In: *Proceedings of the National Academy of Sciences* 110.45 (2013), pp. 18327–18332. DOI: 10.1073/pnas.1306572110.

[26] H Von Helmholtz. 'Physiological Optics'. In: *Uspekhi Fizicheskikh Nauk* III.10 (1925), pp. 1193–1213. DOI: 10.1007/978-3-540-39053-4.

[27] Daniel Kersten and Alan Yuille. 'Bayesian models of object perception'. In: *Current Opinion in Neurobiology* 13.2 (Apr. 2003), pp. 150–158. DOI: 10.1016/S0959-4388(03)00042-4.

[28] Steven Pinker. *The Language Instinct: How the Mind Creates Language*. en. HarperCollins, 1994.

[29] Matthew M. Hurley et al. *Inside Jokes: Using Humor to Reverse-engineer the Mind*. en. MIT Press, 2011.

[30] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998. DOI: 10.1016/S0065-230X(09)04001-9.

[31] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 'Action Understanding as Inverse Planning'. In: *Cognition* 113.3 (Dec. 2009), pp. 329–49. DOI: 10.1016/j.cognition.2009.07.005.

[32] David Hume. *An inquiry concerning human understanding*. en. 1748.

[33] Karl Raimund Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. en. Psychology Press, 2002.

[34] Joshua B Tenenbaum et al. 'How to grow a mind: statistics, structure, and abstraction.' In: *Science (New York, N.Y.)* 331.6022 (Mar. 2011), pp. 1279–85. DOI: 10.1126/science.1192788.

[35] Charles Kemp and Joshua B Tenenbaum. 'The discovery of structural form.' In: *Proceedings of the National Academy of Sciences of the United States of America* 105.31 (2008), pp. 10687–92. DOI: 10.1073/pnas.0802631105.

[36] Roger B Grosse et al. 'Exploiting compositionality to explore a large space of model structures'. In: *Conference on Uncertainty in Artificial Intelligence* (2012).

[37] Steven T Piantadosi, Joshua B Tenenbaum, and Noah D. Goodman. 'Bootstrapping in a language of thought: a formal model of numerical concept learning.' In: *Cognition* 123.2 (May 2012), pp. 199–217. DOI: 10.1016/j.cognition.2011.11.005.

[38] Kevin Ellis et al. *DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning*. June 2020. DOI: 10.48550/arXiv.2006.08381.

[39] R. J. Solomonoff. 'A formal theory of inductive inference. Part I'. In: *Information and Control* 7.1 (1964), pp. 1–22. DOI: 10.1016/S0019-9958(64)90223-2.

[40] Joshua S Rule, Joshua B Tenenbaum, and Steven T Piantadosi. 'The Child as Hacker.' In: *Trends in cognitive sciences* (Oct. 2020). DOI: 10.1016/j.tics.2020.07.005.

[41] David J C MacKay. *Information Theory, Inference, and Learning Algorithms David J.C. MacKay*. Vol. 100. 2005. DOI: 10.1198/jasa.2005.s54.

[42] Freeman Dyson. 'A meeting with Enrico Fermi'. en. In: *Nature* 427.6972 (Jan. 2004), pp. 297–297. DOI: 10.1038/427297a.

[43] Carl Edward Rasmussen and Zoubin Ghahramani. 'Occam's Razor'. In: *Advances in Neural Information Processing* (2000).

[44] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. 'Novel approach to nonlinear/non-Gaussian Bayesian state estimation'. en. In: *IEE Proceedings F (Radar and Signal Processing)* 140.2 (Apr. 1993), pp. 107–113. DOI: 10.1049/ip-f-2.1993.0015.

[45] N. Kantas et al. 'An Overview of Sequential Monte Carlo Methods for Parameter Estimation in General State-Space Models'. en. In: *IFAC Proceedings Volumes*. 15th IFAC Symposium on System Identification 42.10 (Jan. 2009), pp. 774–785. DOI: 10.3182/20090706-3-FR-2004.00129.

[46] George Casella and Christian P. Robert. 'Rao-Blackwellisation of Sampling Schemes'. In: *Biometrika* 83.1 (1996), pp. 81–94.

[47] Ferenc Huszár. *Choice of Recognition Models in VAEs: a regularisation view*. en. Mar. 2017.

[48] S. Gershman and Noah D. Goodman. 'Amortized Inference in Probabilistic Reasoning'. In: *Cognitive Science* (2014).

[49] Ishita Dasgupta and Samuel J. Gershman. 'Memory as a Computational Resource'. en. In: *Trends in Cognitive Sciences* 25.3 (Mar. 2021), pp. 240–251. DOI: 10.1016/j.tics.2020.12.008.

[50] CE Shannon. 'A mathematical theory of communication'. In: *ACM SIGMOBILE Mobile Computing and ...* 27.July 1928 (1948), pp. 379–423.

[51] Naftali Tishby, Fernando C. Pereira, and William Bialek. 'The information bottleneck method'. In: *Proc. of the 37th Allerton Conference on Communication, Control and Computing* (1999).

[52] William Bialek, Ilya Nemenman, and Naftali Tishby. *Predictability, Complexity, and Learning*. Tech. rep. 2001.

[53] Alan D. Baddeley et al. *Memory*. en. Psychology Press, 2009.

[54] David G Nagy and Gergo Orban. 'Episodic memory as a prerequisite for online updates of model structure'. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. 2016, pp. 1–7.

[55] David G. Nagy, Balázs Török, and Gergő Orbán. 'Optimal forgetting: Semantic compression of episodic memories'. In: *PLOS Computational Biology* 16.10 (2020), e1008367. DOI: 10.1371/journal.pcbi.1008367.

[56] David G Nagy, Balazs Torok, and Gergo Orban. 'Rate distortion trade-off in human memory'. In: 2019. DOI: 10.32470/ccn.2019.1115-0.

[57] Csenge Frater, David G. Nagy, and Gergő Orbán. 'Forgetting in delayed recognition as generative compression with decreasing capacity'. en. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 44.44 (2022).

[58] Mihály Bányai, Dávid G. Nagy, and Gergő Orbán. 'Hierarchical semantic compression predicts texture selectivity in early vision'. en. In: *2019 Conference on Cognitive Computational Neuroscience*. Berlin, Germany: Cognitive Computational Neuroscience, 2019. DOI: 10.32470/CCN.2019.1092-0.

[59] Timo Flesch et al. 'Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals'. en. In: *PLOS Computational Biology* 19.1 (Jan. 2023), e1010808. DOI: 10.1371/journal.pcbi.1010808.

[60] James L. McClelland, B L McNaughton, and Randall C. O'Reilly. 'Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory.' In: *Psychological review* 102.3 (July 1995), pp. 419–57.

[61] Máté Lengyel and P Dayan. 'Hippocampal contributions to control: The third way'. In: *Advances in neural information processing systems* (2009), pp. 1–8.

[62] Johannes B. Mahr and Gergely Csibra. 'Why do we remember? The communicative function of episodic memory'. In: *Behavioral and Brain Sciences* 41 (2017), pp. 1–93. DOI: 10.1017/S0140525X17000012.

[63] Masa-aki Sato. 'Online Model Selection Based on the Variational Bayes'. In: *Neural Computation* 13.7 (2001), pp. 1649–1681. DOI: 10.1162/089976601750265045.

[64] Paul Fearnhead. 'Particle filters for mixture models with an unknown number of components'. In: *Statistics and Computing* 14.1 (2004), pp. 11–21. DOI: 10.1023/B:STCO.0000009418.04621.cd.

[65] R. Gomes, M. Welling, and P. Perona. 'Incremental learning of nonparametric Bayesian mixture models'. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008). DOI: 10.1109/CVPR.2008.4587370.

[66] Zoubin Ghahramani. 'Probabilistic machine learning and artificial intelligence'. In: *Nature* 521.7553 (2015), pp. 452–459. DOI: 10.1038/nature14541.

[67] Gergo Orbán et al. 'Bayesian learning of visual chunks by human observers.' In: *Proceedings of the National Academy of Sciences of the United States of America* 105.7 (Feb. 2008), pp. 2745–50. DOI: 10.1073/pnas.0708424105.

[68] Adam N Sanborn, Thomas L. Griffiths, and Daniel J. Navarro. 'A More Rational Model of Categorization'. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (2006), pp. 1–6.

[69] Robert M. French. 'Catastrophic forgetting in connectionist networks'. In: *Trends in Cognitive Sciences* 3.4 (1999), pp. 128–135. DOI: 10.1016/S1364-6613(99)01294-2.

[70] Neil R. Bramley et al. *Formalizing Neurath's Ship: Approximate Algorithms for Online Causal Learning*. Vol. 124. 2016. DOI: 10.1037/rev0000061.

[71] Edward Snelson and Zoubin Ghahramani. 'Compact approximations to Bayesian predictive distributions'. In: *... of the 22nd international conference on ...* Mcmc (2005).

[72] Laurent Itti and Pierre Baldi. 'Bayesian Surprise Attracts Human Attention'. In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Vol. 18. MIT Press, 2005.

[73] Michael McCloskey and Neal J. Cohen. 'Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem'. In: *Psychology of Learning and Motivation - Advances in Research and Theory* 24.C (1989), pp. 109–165. DOI: 10.1016/S0079-7421(08)60536-8.

[74] Timo Flesch et al. 'Comparing continual task learning in minds and machines'. In: *Proceedings of the National Academy of Sciences* 115.44 (2018), p. 201800755. DOI: 10.1073/pnas.1800755115.

[75] Raymond S. Nickerson and Marilyn Jager Adams. 'Long-term memory for a common object'. In: *Cognitive Psychology* 11.3 (1979), pp. 287–307. DOI: 10.1016/0010-0285(79)90013-6.

[76] Maryanne Martin and Gregory V Jones. 'Generalizing everyday memory: Signs and handedness'. In: *Memory and Cognition* 26.2 (1998), pp. 193–200. DOI: 10.3758/BF03201132.

[77] Adam B. Blake, Meenely Nazarian, and Alan D. Castel. 'The apple of the mind's eye: Everyday attention, metamemory, and reconstructive memory for the apple logo'. In: *Quarterly Journal of Experimental Psychology* 68.5 (2015), pp. 858–865. DOI: 10.1080/17470218.2014.1002798.

[78] Frederic Charles Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

[79] Gordon H Bower, Martin B Karlin, and Alvin Dueck. 'Comprehension and memory for pictures'. In: *Memory & Cognition* 3.2 (1975), pp. 216–220. DOI: 10.3758/BF03212900.

[80] L. Carmichael, H. P. Hogan, and A. A. Walter. 'An experimental study of the effect of language on the reproduction of visually perceived forms'. In: *Journal of Experimental Psycology* 15.1 (1932), pp. 73–86. DOI: 10.1037/h0072671.

[81] John D. Bransford and Marcia K. Johnson. 'Contextual prerequisites for understanding: Some investigations of comprehension and recall'. In: *Journal of Verbal Learning and Verbal Behavior* 11.6 (1972), pp. 717–726. DOI: 10.1016/S0022-5371(72)80006-9.

[82] Elizabeth F. Loftus. 'Planting misinformation in the human mind: A 30-year investigation of the malleability of memory'. In: *Learning and Memory* 12.4 (July 2005), pp. 361–366. DOI: 10.1101/lm.94705.

[83] Maryanne Garry et al. 'Imagination inflation: Imagining a childhood event inflates confidence that it occurred'. In: *Psychonomic Bulletin and Review* 3.2 (1996), pp. 208–214. DOI: 10.3758/BF03212420.

[84] Neal J. Roese and Kathleen D. Vohs. 'Hindsight Bias'. In: *Perspectives on Psychological Science* 7.5 (2012), pp. 411–426. DOI: 10.1177/1745691612454303.

[85] Elizabeth F. Loftus and John C. Palmer. 'Reconstruction of automobile destruction: An example of the interaction between language and memory'. In: *Journal of Verbal Learning and Verbal Behavior* 13.5 (1974), pp. 585–589. DOI: 10.1016/S0022-5371(74)80011-3.

[86] JR Anderson and R Milson. 'Human Memory: An Adaptive Perspective'. In: *Psychological Review* (1989).

[87] Daniel L Schacter, Scott a Guerin, and Peggy L St Jacques. 'Memory distortion: an adaptive perspective.' In: *Trends in cognitive sciences* 15.10 (Oct. 2011), pp. 467–74. DOI: 10.1016/j.tics.2011.08.004.

[88] Moshe Bar. 'The proactive brain: using analogies and associations to generate predictions'. In: *Trends in Cognitive Sciences* 11.7 (2007), pp. 280–289. DOI: 10.1016/j.tics.2007.05.005.

[89] Eryn J. Newman and D. Stephen Lindsay. 'False memories: What the hell are they for?' In: *Applied Cognitive Psychology* 23.8 (2009), pp. 1105–1121. DOI: 10.1002/acp.1613.

[90] Janellen Huttenlocher, Larry V Hedges, and Susan Duncan. 'Categories and particulars: Prototype effects in estimating spatial location.' In: *Psychological Review* 98.3 (1991), pp. 352–376. DOI: 10.1037/0033-295X.98.3.352.

[91] Timothy Brady, Daniel Schacter, and George Alvarez. 'The Adaptive Nature of False Memories is Revealed by Gist-based Distortion of True Memories'. In: *PsyArXiv* (2018). DOI: 10.31234/osf.io/zeg95.

[92] Alan Yuille and Daniel Kersten. 'Vision as Bayesian inference: analysis by synthesis?' In: *Trends in cognitive sciences* 10.7 (July 2006), pp. 301–8. DOI: 10.1016/j.tics.2006.05.002.

[93] Diederik P Kingma and Max Welling. 'Auto-Encoding Variational Bayes'. In: *International Conference on Learning Representations* (2013). DOI: 10.1051/0004-6361/201527329.

[94] Troy Chinen et al. 'Towards a Semantic Perceptual Image Metric'. In: *25th IEEE International Conference on Image Processing* (2018), pp. 1–5.

[95] Shibani Santurkar, David Budden, and Nir Shavit. 'Generative Compression'. In: *2018 Picture Coding Symposium (PCS)* (2017).

[96] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. 'Bayesian theories of conditioning in a changing world'. In: *Trends in Cognitive Sciences* 10.7 (2006), pp. 294–300. DOI: 10.1016/j.tics.2006.05.004.

[97] David M. Sobel, Joshua B. Tenenbaum, and Alison Gopnik. 'Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers'. In: *Cognitive Science* 28.3 (2004), pp. 303–333. DOI: 10.1016/j.cogsci.2003.11.001.

[98] Samuel J. Gershman, Kenneth a. Norman, and Yael Niv. 'Discovering latent causes in reinforcement learning'. In: *Current Opinion in Behavioral Sciences* (2015), pp. 1–8. DOI: 10.1016/j.cobeha.2015.07.007.

[99] Alexander A. Alemi et al. 'Deep Variational Information Bottleneck'. In: *International Conference on Learning Representations* (2016).

[100] Alexander A. Alemi et al. 'Fixing a Broken ELBO'. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2017).

[101] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. 'End-to-end Optimized Image Compression'. In: *International Conference on Learning Representations* (2016). DOI: 10.1016/S0197-3975(03)00059-6.

[102] Irina Higgins et al. 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *International Conference on Learning Representations* (2017).

[103] Fernand Gobet and Herbert A. Simon. 'Recall of rapidly presented random chess positions is a function of skill'. In: *Psychonomic Bulletin and Review* 3.2 (1996), pp. 159–163. DOI: 10.3758/BF03212414.

[104] A. D. Baddeley. 'Immediate memory and the "perception" of letter sequences'. In: *Quarterly Journal of Experimental Psychology* 16.4 (Dec. 1964), pp. 364–367. DOI: 10.1080/17470216408416394.

[105] A D Baddeley. 'Language Habits, Acoustic Confusability, and Immediate Memory for Redundant Letter Sequences'. In: *Psychonomic Science* 22.2 (1971), pp. 120–121. DOI: 10.3758/BF03332525.

[106] Alec Radford et al. 'Language Models are Unsupervised Multitask Learners'. In: *OpenAI blog* (2018).

[107] Henry L. Roediger and Kathleen B. McDermott. 'Creating False Memories: Remembering Words Not Presented in Lists'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21.4 (1995), pp. 803–814. DOI: 10.1037/0278-7393.21.4.803.

[108] Yishu Miao, Lei Yu, and Phil Blunsom. 'Neural Variational Inference for Text Processing'. In: *Proc. of the 33rd International Conference on International Conference on Machine Learning* 48.Mcmc (2016). DOI: 10.1055/s-2008-1070477.

[109] David Ha and Douglas Eck. 'A Neural Representation of Sketch Drawings'. In: *International Conference on Learning Representations* (2017), pp. 1–20.

[110] N. G. Hanawalt and I. H. Demarest. 'The effect of verbal suggestion upon the reproduction of visually perceived forms'. In: *Journal of Experimental Psychology* 25.2 (1939), pp. 159–174. DOI: 10.1037/h0057682.

[111] John R. Anderson and Lael J. Schooler. 'Reflections of the Environment in Memory'. In: *Psychological Science* 2.6 (1991), pp. 396–408. DOI: 10.1111/j.1467-9280.1991.tb00174.x.

[112] H. L. Roediger and K. A. DeSoto. 'Forgetting the presidents'. In: *Science* 346.6213 (Nov. 2014), pp. 1106–1109. DOI: 10.1126/science.1259627.

[113] Michael P. Toglia, Jeffrey S. Neuschatz, and Kerri A. Goodwin. 'Recall Accuracy and Illusory Memories: When More is Less'. In: *Memory* 7.2 (1999), pp. 233–256. DOI: 10.1080/741944069.

[114] John G Seamon et al. 'Are false memories more difficult to forget than accurate memories? The effect of retention interval on recall and recognition'. In: *Memory and Cognition* 30.7 (2002), pp. 1054–1064. DOI: 10.3758/BF03194323.

[115] Anjali Thapar and Kathleen B. McDermott. 'False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing'. In: *Memory and Cognition* 29.3 (2001), pp. 424–432. DOI: 10.3758/BF03196393.

[116] Kim J. Vicente and Jo Anne H. Wang. 'An Ecological Theory of Expertise Effects in Memory Recall'. In: *Psychological Review* 105.1 (1998), pp. 33–57. DOI: 10.1037/0033-295X.105.1.33.

[117] Josh H. McDermott, Michael Schemitsch, and Eero P. Simoncelli. 'Summary statistics in auditory perception'. In: *Nature Neuroscience* 16.4 (2013), 493–U169. DOI: 10.1038/nn.3347.

[118] Christopher P Burgess et al. 'Understanding disentangling in beta-VAE'. In: *Advances in Neural Information Processing* Nips (2017).

[119] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. 'Variable rate deep image compression with a conditional autoencoder'. In: *Proceedings of the IEEE International Conference on Computer Vision* 2019-Octob (2019), pp. 3146–3154. DOI: 10.1109/ICCV.2019.00324.

[120] Yibo Yang, Robert Bamler, and Stephan Mandt. 'Variable-Bitrate Neural Compression via Bayesian Arithmetic Coding'. In: *arXiv:2002.08158* (2020).

[121] Karol Gregor et al. 'Towards Conceptual Compression'. In: *Arxiv* (2016), p. 14.

[122] Lars Maaløe et al. 'BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling'. In: *Advances in Neural Information Processing* (2019).

[123] Tero Karras et al. 'Progressive growing of GANs for improved quality, stability, and variation'. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. 2018, pp. 1–26.

[124] TS Lee and David Mumford. 'Hierarchical Bayesian inference in the visual cortex.' In: *Journal of the Optical Society of America. A, Optics, image science, and vision* 20.7 (2003), pp. 1434–48. DOI: 10.1364/JOSAA.20.001434.

[125] Corey M. Ziemba et al. 'Selectivity and tolerance for visual texture in macaque V2'. In: *Proceedings of the National Academy of Sciences* 113.22 (2016), E3140–E3149. DOI: 10.1073/pnas.1510847113.

[126] Samuel J Gershman. 'Predicting the past, remembering the future'. In: *Current Opinion in Behavioral Sciences* 17 (2017), pp. 7–13. DOI: 10.1016/j.cobeha.2017.05.025.

[127] Cuong V Nguyen et al. 'Variational continual learning'. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. 2018.

[128] DJ Strouse and David J Schwab. 'The deterministic information bottleneck'. In: *Arxiv* 1991 (2016), p. 15.

[129] Christopher J. Bates and Robert A. Jacobs. 'Efficient Data Compression in Perception and Perceptual Memory'. In: *Psychological Review* (2020). DOI: 10.1037/rev0000197.

[130] David G. Nagy, Balázs Török, and Gergő Orbán. 'Semantic Compression of Episodic Memories'. In: *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. 2018.

[131] Danilo Jimenez Rezende and Fabio Viola. 'Taming VAEs'. In: *Arxiv* (2018).

[132] Yochai Blau and Tomer Michaeli. 'Rethinking lossy compression: The rate-distortion-perception tradeoff'. In: *36th International Conference on Machine Learning, ICML 2019* 2019-June (2019), pp. 1081–1091.

[133] Ian Goodfellow et al. 'Generative Adversarial Nets'. In: *Advances in Neural Information Processing Systems 27* (2014), pp. 2672–2680.

[134] Chris R Sims, Robert A Jacobs, and David C Knill. 'Supplementary Material to: An ideal observer analysis of visual working memory.' In: *Psychological review* 119.4 (2012), pp. 807–30. DOI: 10.1037/a0029856.

[135] Chris R Sims. 'The cost of misremembering: Inferring the loss function in visual working memory'. In: *Journal of Vision* 15.3 (2015), pp. 1–27. DOI: 10.1167/15.3.2.doi.

[136] Chris R Sims et al. 'Exploring the Cost Function in Color Perception and Memory: An Information-Theoretic Model of Categorical Effects in Color Matching'. In: *CogSci* (2016), pp. 2273–2278.

[137] Jon W. Carr et al. 'Simplicity and informativeness in semantic category systems'. In: *Cognition* 202.July 2018 (2020), p. 104289. DOI: 10.1016/j.cognition.2020.104289.

[138] Talya Sadeh et al. 'How we forget may depend on how we remember'. In: *Trends in Cognitive Sciences* 18.1 (2014), pp. 26–36. DOI: 10.1016/j.tics.2013.10.008.

[139] Dharshan Kumaran, Demis Hassabis, and James L. McClelland. 'What learning systems do intelligent agents need? Complementary Learning Systems Theory Updated'. In: *Trends in Cognitive Sciences* 20.7 (2016), pp. 512–534. DOI: 10.1016/j.tics.2016.05.004.

[140] Greg Wayne et al. 'Unsupervised Predictive Memory in a Goal-Directed Agent'. In: *arXiv preprint* (2018).

[141] Alexander Pritzel and Benigno Uria. 'Neural Episodic Control'. In: *Proceedings of the 34th International Conference on Machine Learning* (2017).

[142] Wickliffe C. Abraham and Anthony Robins. 'Memory retention - The synaptic stability versus plasticity dilemma'. In: *Trends in Neurosciences* 28.2 (2005), pp. 73–78. DOI: 10.1016/j.tins.2004.12.003.

[143] Blake A. Richards and Paul W. Frankland. 'The Persistence and Transience of Memory'. In: *Neuron* 94.6 (2017), pp. 1071–1084. DOI: 10.1016/j.neuron.2017.04.037.

[144] Adam Santoro, Paul W. Frankland, and Blake A. Richards. 'Memory transformation enhances reinforcement learning in dynamic environments'. In: *Journal of Neuroscience* 36.48 (2016), pp. 12228–12242. DOI: 10.1523/JNEUROSCI.0763-16.2016.

[145] Konrad P. Kording, Joshua B. Tenenbaum, and Reza Shadmehr. 'The dynamics of memory as a consequence of optimal adaptation to a changing body'. In: *Nature Neuroscience* 10.6 (2007), pp. 779–786. DOI: 10.1038/nn1901.

[146] Oliver Hardt, Einar On Einarsson, and Karim Nader. 'A Bridge Over Troubled Water: Reconsolidation as a Link Between Cognitive and Neuroscientific Memory Research Traditions'. In: *Annual Review of Psychology* 61.1 (2010), pp. 141–167. DOI: 10.1146/annurev.psych.093008.100455.

[147] Mark Steyvers and Thomas L. Griffiths. 'Rational analysis as a link between human memory and information retrieval'. In: *The Probabilistic Mind* (2008).

[148] David G. Nagy, Balázs Török, and Gergő Orbán. 'Semantic Compression of Episodic Memories'. In: *Proc. of Conference on Cognitive Computational Neuroscience*. 2019. DOI: 10.32470/ccn.2018.1050-0.

[149] Samuel J Gershman et al. 'The computational nature of memory modification'. In: *eLife* 6 (Mar. 2017), pp. 1–40. DOI: 10.7554/eLife.23763.

[150] Samuel J. Gershman et al. 'Statistical Computations Underlying the Dynamics of Memory Updating'. In: *PLoS Computational Biology* 10.11 (Nov. 2014). Ed. by Olaf Sporns, e1003939. DOI: 10.1371/journal.pcbi.1003939.

[151] Jeremy Freeman et al. 'A functional and perceptual signature of the second visual area in primates'. en. In: *Nature Neuroscience* 16.7 (July 2013), pp. 974–981. DOI: 10.1038/nn.3402.

[152] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. 'How Does the Brain Solve Visual Object Recognition?' en. In: *Neuron* 73.3 (Feb. 2012), pp. 415–434. DOI: 10.1016/j.neuron.2012.01.010.

[153] Daniel L. K. Yamins et al. 'Performance-optimized hierarchical models predict neural responses in higher visual cortex'. In: *Proceedings of the National Academy of Sciences* 111.23 (June 2014), pp. 8619–8624. DOI: 10.1073/pnas.1403112111.

[154] Santiago A. Cadena et al. 'Deep convolutional models improve predictions of macaque V1 responses to natural images'. en. In: *PLOS Computational Biology* 15.4 (Apr. 2019), e1006897. DOI: 10.1371/journal.pcbi.1006897.

[155] Kohitij Kar et al. 'Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior'. en. In: *Nature Neuroscience* 22.6 (June 2019), pp. 974–983. DOI: 10.1038/s41593-019-0392-5.

[156] Tai Sing Lee and My Nguyen. 'Dynamics of subjective contour formation in the early visual cortex'. In: *Proceedings of the National Academy of Sciences* 98.4 (Feb. 2001), pp. 1907–1911. DOI: 10.1073/pnas.98.4.1907.

[157] Ruben Coen-Cagli, Peter Dayan, and Odelia Schwartz. 'Cortical Surround Interactions and Perceptual Salience via Natural Scene Statistics'. en. In: *PLOS Computational Biology* 8.3 (Mar. 2012), e1002405. DOI: 10.1371/journal.pcbi.1002405.

[158] Gergő Orbán et al. 'Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex'. en. In: *Neuron* 92.2 (Oct. 2016), pp. 530–543. DOI: 10.1016/j.neuron.2016.09.038.

[159] Mihály Bányai et al. 'Stimulus complexity shapes response correlations in primary visual cortex'. In: *Proceedings of the National Academy of Sciences* 116.7 (Feb. 2019), pp. 2723–2732. DOI: 10.1073/pnas.1816766116.

[160] BA Olshausen. 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images'. In: *Nature* (1996).

[161] Gabriel Barello, Adam S. Charles, and Jonathan W. Pillow. *Sparse-Coding Variational Auto-Encoders*. en. Aug. 2018. DOI: 10.1101/399246.

[162] Javier Portilla and Eero P. Simoncelli. 'A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients'. en. In: *International Journal of Computer Vision* 40.1 (Oct. 2000), pp. 49–70. DOI: 10.1023/A:1026553619983.

[163] Peter Dayan et al. 'The Helmholtz Machine'. In: *Neural Computation* 7.5 (Sept. 1995), pp. 889–904. DOI: 10.1162/neco.1995.7.5.889.

[164] Emily Fertig et al. 'beta-VAEs can retain label information even at high compression'. In: NeurIPS (2018).

[165] Casper Kaae Sønderby et al. 'Ladder Variational Autoencoders'. In: (2016).

[166] J. H. van Hateren and A. van der Schaaf. 'Independent component filters of natural images compared with simple cells in primary visual cortex'. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1394 (Mar. 1998), pp. 359–366. DOI: 10.1098/rspb.1998.0303.

[167] Erkki Oja. 'Principal components, minor components, and linear neural networks'. In: *Neural Networks* 5.6 (1992), pp. 927–935. DOI: 10.1016/S0893-6080(05)80089-9.

[168] P. Foldiak. 'Learning Invariance from Transformation Sequences'. In: *Neural Computation* 3 (1991), p. 2853.

[169] William G. Chase and Herbert A. Simon. 'Perception in chess'. In: *Cognitive Psychology* 4.1 (1973), pp. 55–81. DOI: 10.1016/0010-0285(73)90004-2.

Permissions and copyright

1. Necker cube & impossible cube from Wikimedia Commons, by User:Boivie and User:Maksim.

2. All M.C. Escher works © 2023 The M.C. Escher Company - the Netherlands. All rights reserved. Used with permission. www.mcescher.com

3. Blue eyed girl colour constancy illusion created by Akiyoshi Kitaoka. https://www.ritsumei.ac.jp/ akitaoka/index-e.html

4. Fig. 1.13 A, C reprinted from Grosse et al. [36], B, D reprinted from Kemp et al. [35].

5. Fig. 4.7 reprinted from Flesch & Nagy et al. [59]. Used with permission.

6. Vector graphics from The Noun Project, by Max Hancock, Muneer A. Safiah, Stefania Servidio, Maria Lora Macias, Creative Stall, Nikita Kozin, Hugo Alberto, Joe Pictos, Sumana Chamrunworakiat, Dmitry Kovalev, Javier Cabezas and André Luiz Gollo.

Thesis summary

Learning a simplified internal model of the environment and using it for prediction, perception and decision making is a crucial component of the human brain's ability to adapt to environmental demands. Maintaining general world knowledge in the form of a probabilistic generative model of the environment has been identified as the main goal of semantic memory, one of the two complementary memory systems comprising long-term memory in humans. In contrast, the second memory system called episodic memory aims to retain a detailed representation of environmental variables during specific events. This thesis aims to provide a computation level account of long-term memory in humans and other intelligent agents operating under similar resource constraints, focusing on the interactions of the semantic memory and episodic memory systems.

In the first study, I aim to understand why an episodic memory system is a necessary ingredient for learning agents and propose that it is required to resolve a fundamental challenge of online model selection under resource constraints.

In a second study, I explore how the probabilistic generative model maintained in semantic memory can support the episodic memory system. I propose that utilising the probabilistic model for generative compression, a recently developed approach in the field of machine learning, enables efficient storage of episodic memories. Furthermore, I demonstrate that such *semantic compression* parsimoniously explains a wide range of memory distortions that have been observed in the experimental literature on human memory.

In the final two studies I explore how some of the above computational level ideas can be implemented in neural networks. In the first, we propose that implementing semantic compression with hierarchical generative models can explain features of multi-unit recordings of neural activity in the visual cortex of macaques. In the second, I propose a scheme for how partitioned task representations can be stored in artificial neural networks, enabling maintained accuracy in the setting known as continual learning in the machine learning literature as well as applying this model for explaining puzzling features of a behavioural experiment in humans.

Összefoglaló

A környezet egyszerűsített belső modelljének elsajátítása, illetve predikcióban, észlelésben és döntéshozatalban való felhasználása az emberi agy környezeti igényekhez való alkalmazkodási képességének alapvető komponensei. A világról való általános ismeretek a környezet valószínűségi generatív modelljének formájában való fenntartása a szemantikus memória fő célja, amely az emberi hosszú távú memóriát alkotó két memóriarendszer egyike. A másik rendszer, az epizodikus memória, a környezeti változók részletes reprezentációjának megtartását célozza. A disszertáció célja, hogy az emberek és más, hasonló erőforrás-korlátozások mellett működő intelligens ágensek memóriarendszereinek komputációs szintű leírását adja, a szemantikus memória és az epizodikus memóriarendszerek kölcsönhatásaira összpontosítva.

Az első tanulmányban arra törekszem, hogy megértsem, miért szükséges az epizodikus memóriarendszer a tanulási ágensek számára, és azt vetem fel, hogy nélkülözhetetlen az online modellválasztás egy alapvető kihívásának erőforráskorlátok melletti megoldásához.

A második tanulmányban azt vizsgálom, hogy a szemantikus memóriában fenntartott valószínűségi generatív modell hogyan tudja támogatni az epizodikus memóriarendszert. Megmutatom, hogy a valószínűségi modell felhasználása az epi-zódok generatív tömörítésére, amely egy a gépi tanulás területén nemrégiben kifejlesztett megközelítés, lehetővé teszi az epizodikus emlékek hatékony tárolását. Továbbá demonstrálom, hogy ez az ú.n. *szemantikus tömörítés* egységesítő magyarázatként szolgál az emberi memória kísérleti irodalmában megfigyelt szisztematikus memóriatorzulások széles skálájára.

Az utolsó két tanulmányban azt vizsgáljuk, hogy a fenti számítási szintű ötletek némelyike hogyan valósítható meg neurális hálózatokban. Az elsőben megmutatjuk, hogy a szemantikus tömörítés hierarchikus generatív modellekkel való megvalósítása magyarázni tudja makákók látókérgében mért neurális aktivitás egyes jellemzőit. A másodikban egy sémát javasolok feladatok reprezentációinak mesterséges neurális hálózatokban való struktúrált tárolására, lehetővé téve a gépi tanulás irodalomban folyamatos tanulásként ismert probléma feloldására egy konkrét kísérleti környezetben, amelyet felhasználunk ugyanebben a kísérletben az emberi tanulás jellemzőinek modellezésére.