Precis of *A Bayesian account of learning algorithms and generalising representations in the brain*

James C.R. Whittington

Chapter 1: Introduction

Without learning we would be limited to a set of preprogrammed behaviours. While that may be acceptable for flies[1], it does not provide the basis for adaptive or intelligent behaviours familiar to humans. Learning, then, is one of the crucial components of brain operation. Learning, however, takes time. Thus, the key to adaptive behaviour is learning to systematically generalise; that is, have learned knowledge that can be flexibly recombined to understand any world in front of you. This thesis attempts to make inroads on two questions - how can brain networks learn, and what are the principles behind representations of knowledge that allow generalisation.

With the industrialisation of science, the twentieth century bore fruit in the form of an increasingly detailed understanding of neurons, synapses, neurotransmitters, resting potentials, action potentials, networks and so on (1–4). Though we have gained a great level of detail about many of these micro-processes - as well as high-level understandings of intelligence thanks to philosophy, experimental psychology, and behavioural and cognitive neuroscience (5–9) - a large gulf of understanding remains between these levels of granularity.

This thesis focuses on spanning this gap by providing high-level computational frameworks that translate to low-level processes. Any high-level brain framework must have successful behaviour at its heart as that is the role of the brain. Analogously, neurons are central to low-level understanding as the basis of brain function is believed to be the transfer of information between neurons, mediated via weighted connections. Different weights lead to different functions. Thus, learning appropriate configurations of weights is the fundamental problem facing brains. There are two facets to this learning - the first is *how*, and the second is *what*. The how are the learning *algorithms* that determine updates to these synaptic connections, and the what are the neural *representations* that reflect how the world works.

In this vein, this thesis examines 1) the algorithmic implementation of learning in biological neural networks, and 2) a computational framework for the neural representations of task generalisation. Both these research directions are bound together by Bayesian thinking, and both of these pieces of work bridge the gap between high- and low- level understanding, as well as between brains and machines.

Artificial and brain networks

Learning is the process of updating one's belief or understanding on the basis of new information. This is a problem for machines and, more intimately, it is the problem faced by the neurons and connections

[1] Mayflies to be more specific!

that comprise our brains. Machine learners get to choose their learning algorithm. One algorithm in particular, the back-propagation algorithm (backprop; (10; 11)), is proven to be highly effective.

Though ANNs are loosely based on brain networks, backprop is traditionally seen as incompatible with learning in the brain. The main contribution of the first part of this thesis is to show that the backprop is, in fact, implementable in locally connected networks of neurons such as those found in the brain.

We verify this equivalence on a standard machine learning benchmark, demonstrating equal performance for both algorithms. Finally, we propose alternative architectures implementing efficient learning in biological neural networks.

After introducing this new network, we compare a variety of other models of biological backprop and demonstrate that many models can be viewed under a common framework.

Neural representations for generalisation

Learning is not done all at once nor is it all for one moment; we learn continually, acquiring and deploying new skills when needed. Learning, however, is slow, and so we learn so that we don't have to learn; we learn so that we can operate successfully in the wild when facing new and/or unknown data.

Applying old knowledge to new situations is known as generalisation. This ranges from classifying novel dogs to complex tasks where entire environments may change. The hippocampal formation has a pivotal role in generalisation, though little is known about the mechanism, or the role of its famous neural representations (12; 13).

The main contribution of the second part of this thesis is threefold. 1) We formalise tasks for structure generalisation - how to transfer information from one task to a structurally similar task. This task setup encompasses seemingly disparate spatial, non-spatial and classic relational memory tasks that all rely on the hippocampal formation. This formalism, then, offers a unifying account of the hippocampal formation. 2) We provide both a computational framework to solve these tasks, and a particular implementation of such ideas - the Tolman-Eichenbaum Machine (TEM). After learning, TEM exhibits neural representations analogous to those found in the hippocampal formation. Thus, TEM provides a mechanistic understanding of the hippocampal role in generalisation, while also offering explanations to the myriad of known cell representations. 3) We verify a neural prediction of this model, showing a previously unknown relationship between cells of hippocampus and entorhinal cortex.

Chapter 2: Bayes and the Brain

Chapter 2 introduces concepts fundamental to the overall research in this thesis. In particular, it introduces the reader to Bayes, sets the scene for algorithmic implementations of Bayes in the brain, and discusses philosophical considerations for structuring representations. There are two particularly relevant sections in this chapter to this thesis:

The first discusses how a method of Bayesian inference, namely variational inference, can translate the problem of inferring a posterior over latent causes to updating neurons in a biologically plausible neural network. Traditionally, these predictive coding networks perform *unsupervised* learning, and have been shown to match various representations in visual cortex. While the overall framework of predictive coding is extremely attractive to neuroscientists, the achievements of predictive coding networks has been limited and its relationship to the hugely successful deep learning is very much unknown.

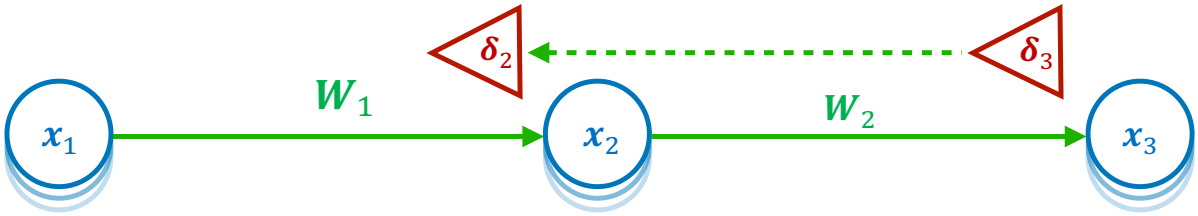

Figure 1: **Artificial neural networks and backprop.** A) Layers of neuron-like nodes are represented by sets of stacked blue circles. Feedforward connections are indicated by green arrows. Backpropagating error signals are shown as red triangles.

The second draws deep parallels between representation learning, and symmetry/invariant structure discovery, as symmetries by their nature, generalise to many situations. We consider recent work (14) that proposes different symmetries should be represented using separate neural populations. This work offers a unique perspective on the machine learning concept of factorised (or disentangled) representations, and is particularly relevant to Chapters 7-9, where we posit that representations of structure should be separated from representations of content.

Part I
Algorithms for learning

Chapter 3: Learning in Artificial and Biological Neural Networks

Chapter 3 introduces concepts fundamental to the new research in this part of the thesis. In particular, it focusses on ANNs, and the decades old debate as to whether their learning algorithm, backprop, is implementable in the brain.

In the past few years, computer programs using deep learning have achieved impressive results in complex cognitive tasks that were previously only in the reach of humans (15–17). Since these recent deep learning applications use extended versions of classic artificial neural networks (ANNs; (10)), their success has inspired studies comparing information processing in ANNs and the brain (18–21).

A key question that remains open is how the brain could implement the error back-propagation algorithm used in ANNs. In a conventional ANN, each node receives a weighted sum of all the nodes from the previous layer (Figure 1). The input layer, $\mathbf{x}_1$, is first set to be the input pattern, $\mathbf{s}$, and then a prediction is made by propagating the activity through the layers according to the following

$$\mathbf{x}_l = W_{l-1} f\left(\mathbf{x}_{l-1}\right) + \mathbf{b}_l$$

Where $\mathbf{x}_l$ is a vector denoting neurons in layer l, W_{l-1} is a matrix of synaptic weights from layer $l-1$ to layer l, and $\mathbf{b}_l$ is a bias. An activation function f is applied to each neuron to allow for non-linear computations.[2]

[2] Alternatively the activation function can be applied after the affine transformation.

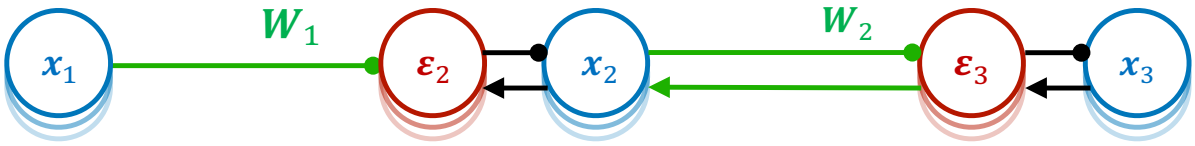

Figure 2: **Predictive coding network architecture.** Arrows and lines ending with circles denote excitatory and inhibitory connections respectively. Connections without labels have weights fixed to 1.

During learning, the synaptic connections are modified to minimise a cost function quantifying the discrepancy between the predicted and target patterns ($\mathbf{t}$), e.g. $E = \frac{1}{2}(\mathbf{t} - \mathbf{x}_L)^T(\mathbf{t} - \mathbf{x}_L)$. The weights are modified in the direction of steepest decrease (or gradient) of the cost function

$$\Delta W_a = \boldsymbol{\delta}_{a+1} f(\mathbf{x}_a)^T$$

With each error ($\boldsymbol{\delta}_a$) getting 'back-propagated' from the error higher up

$$\boldsymbol{\delta}_a = \begin{cases} \mathbf{t} - \mathbf{x}_a & \text{if } a = L \\ \left(W_a^T \boldsymbol{\delta}_{a+1}\right) \odot f'(\mathbf{x}_a) & \text{if } a < L \end{cases}$$

Although the algorithmic process described above appears simple enough, there are serious problems with implementing it in biology. In this chapter we highlight three key issues; 1) Back-prop requires a local error representation, 2) backprop requires symmetry of forwards and backwards weights, and 3) ANNs have unrealistic models of neurons.

Chapter 4: Predictive coding networks implement back-propagation

Chapter 4 shows backprop is implementable in locally connected networks like the brain. In particular it is uses a probabilistic approach to supervised learning, and demonstrates that simple approximate Bayesian inference leads to dynamics that can be implemented in biologically plausible neural networks known as predictive coding networks [3]. The dynamics are

$$\dot{\mathbf{x}}_a = -\boldsymbol{\varepsilon}_a + \left(\Theta_a^T \boldsymbol{\varepsilon}_{a+1}\right) \odot f'(\mathbf{x}_{a+1})$$

Where $\boldsymbol{\varepsilon}_a = \mathbf{x}_a - \boldsymbol{\mu}_a$ are 'error' neurons - neurons that compute the error of the $\mathbf{x}$ neurons compared to what is predicted from the layer below ($\boldsymbol{\mu}_l = \Theta_{l-1} f(\mathbf{x}_{l-1}) + \mathbf{b}_l$). In words, each neuron ($\mathbf{x}_a$) gets is updated according to its prediction error ($\boldsymbol{\varepsilon}_a$), and according to how well it predicts the layer above ($\left(\Theta_a^T \boldsymbol{\varepsilon}_{a+1}\right) \odot f'(\mathbf{x}_{a+1})$). This network of neurons can easily be visualised in Figure 2.

After convergence of the network, weights are updated as

$$\Delta\Theta_a = \boldsymbol{\varepsilon}_{a+1} f(\mathbf{x}_a)^T$$

Importantly, however, when this network reaches dynamic equilibrium, then

[3] This work is published in Neural Computation (22).

$$\varepsilon_a = \begin{cases} \mathbf{t} - \boldsymbol{\mu}_a^* & \text{if } a = L \\ \left(\Theta_a^T \, \varepsilon_{a+1}\right) \odot f'\left(\mathbf{x}_{(a+1)}\right) & \text{if } a < L \end{cases}$$

The errors appear to have back-propagated! These simple equations have exactly the same from as backpropagation and ANNs from Chapter 3, though they are now taking place in locally connected neurons using local learning rules.

The remaining parts of this chapter are devoted to showing the effectiveness of predictive coding networks on a standard machine learning benchmark, as well as describing the conditions under which predictive coding networks and back-propagation become identical.

Chapter 5: Integrating biological models of back-propagation

Chapter 5 reviews other recent models of back-propagation in the brain, summarising them in two categories - those that represent errors explicitly and those that represent errors temporally. It further describes how many models can be viewed as performing inference and learning in an energy-based framework. This work is published in Trends in Cognitive Sciences (23).

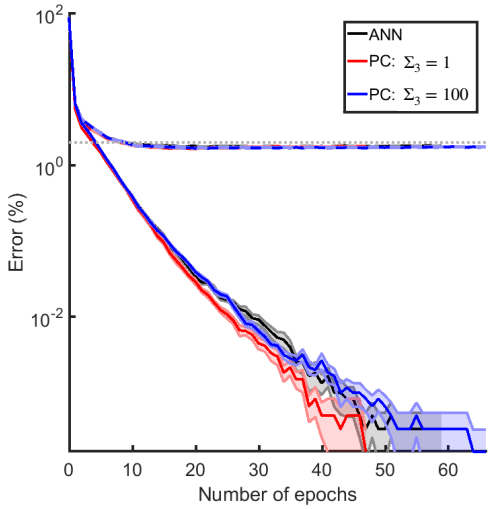

Figure 3: **Predictive coding performance on complex tasks.** Comparison of prediction accuracy (%) for different models (indicated by colours - see key) on the MNIST dataset. Training errors are shown with solid lines, and validation errors with dashed lines. The dotted grey line denotes 2% error. The shaded regions in the fainter colour describe the standard error of the mean. The figure is shown on a logarithmic plot.

Part II
Representations for generalisation

Chapter 6: Generalisation, space and relational memory in the hippocampal formation

Chapter 6 formally introduces generalisation, then swiftly moves on to review the brain's involvement in generalisation - in particular the role and representations of the spatial neurons of the hippocampal formation e.g. place and landmark cells (12; 24) in hippocampus and grid, band object-vector, boundary cells (13; 25–27)) in entorhinal cortex.

This chapter then considers the other function that makes hippocampus so famous; relational memories (58). Relational memories bind together parts of experience, linking memories via their common elements. Relational memories and space have been difficult to square together, with many neuroscientists considering them two very different things.

Chapter 7: The Tolman-Eichenbaum Machine for neuroscientists

Chapter 7 introduces the Tolman-Eichenbaum machine (TEM), a model that learns and generalises abstract relational knowledge in both spatial and non-spatial tasks. The framework of TEM unifies two fields of neuroscience - spatial cognition and relational memory. Furthermore, TEM manages to predict a wide variety of cellular representations.[4]

We account for the broad set of hippocampal properties by re-casting both spatial and relational memory problems as examples of structural abstraction (46) and generalisation. Structural generalisation offers dramatic benefits for new learning and flexible inference, and is a key issue in artificial intelligence. We suggest using "factorised" representations, where which different aspects of knowledge are represented separately and can then be flexibly re-combined to represent novel experiences (47). Factorising the relationships between experiences from the content of each experience could offer a powerful mechanism for generalising this structural knowledge to new situations. Notably, exactly such a factorisation exists between sensory and spatial representations in lateral (LEC) and medial (MEC) entorhinal cortices respectively (48). Manns and Eichenbaum propose that novel conjunctions of these two representations form the hippocampal representation for relational memory.

We demonstrate that this factorisation and conjunction approach is sufficient to build a relational memory system (the Tolman-Eichenbaum machine; TEM) that generalises structural knowledge in space and non-space. The model is a neural network that consisting **only** of two components 1) a recurrent neural network $\mathbf{g}$ that learns and generalises the structure/rules across environments and 2) a Hebbian memory network $\mathbf{p}$ that forms memories for understanding the current environment (particular sensory configuration for each environment).

After formalising the TEM model, this chapter shows that TEM's structural ($\mathbf{g}$) representations resemble grid cells (Figure 4A, 4B for hexagonal and square environments respectively) and band cells (Figure 4D) as recorded in rodent MEC (13; 18; 25; 49). As in the brain, we observe modules of grid

[4] An early version of the work of chapters 7, 8 and 9 is published at NeurIPS (19) and a complete version is published in Cell (45).

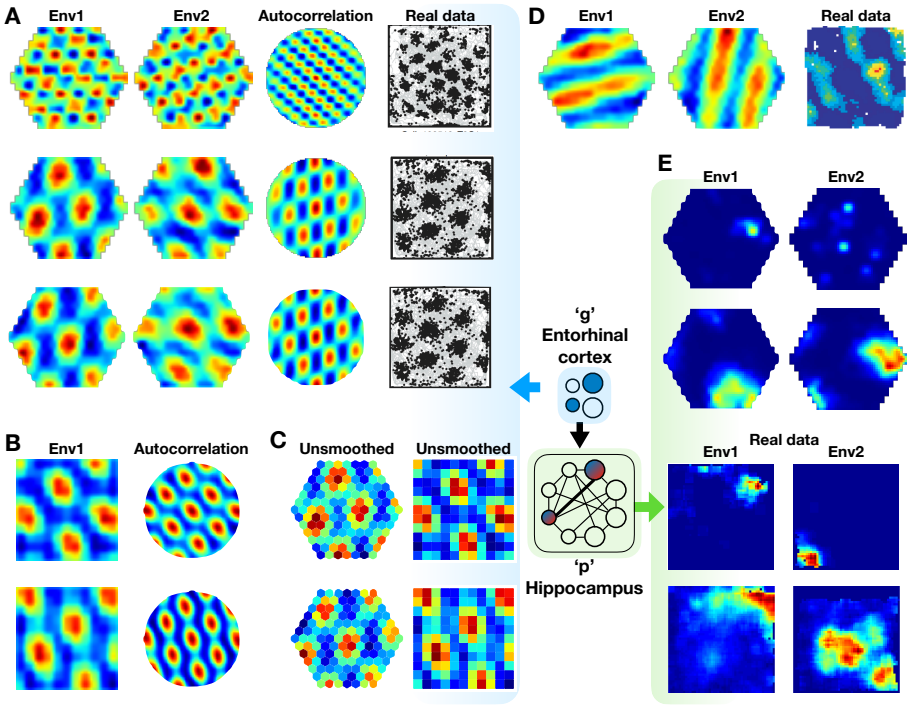

Figure 4: **TEM structural neurons g learn to be grid cells that generalise and TEM conjunctive memory neurons p learn to be place cells that remap.** We use 2D graphs as our environments and a cell's rate map is obtained by allowing the agent to explore the environment then calculating its average firing rate at each point (graph node) in the environment. **A-B)** TEM learned structural representations for random walks on 2D graphs. **A)** Hexagonal worlds. Left to right: environments 1, 2, autocorrelation, real data (25; 65), top to bottom: different cells. TEM learns grid-like cells, of different frequencies (top vs middle), and of different phases (middle vs bottom). **B)** Square worlds. Two TEM learned structural cells - left/right; rate map/ autocorrelation. **C)** Raw unsmoothed rate maps. Left/right: bottom two cells from A/ both cells from B. **D)** TEM also learn band-like cells. Importantly all TEM structural representations A-D) generalise across environments. **E)** Learned memory representations resemble place cells (left/right: environments 1/2; top 2 simulated, bottom 2 real cells) and have different field sizes. These cells remap between environments, i.e. do not generalise.

cells at different spatial frequencies and, within module, we observe cells at different grid phases (Figure 4A).

In TEM, 'hippocampal' cells, $\mathbf{p}$, learns sparse representations that resemble hippocampal place cells (Figure 4E). These place-like fields span multiple sizes, mirroring the hierarchical composition of hippocampal place fields (51; 52). Importantly TEM's 'hippocampal' cells, unlike their 'medial entorhinal' counterparts, do not generalise, but instead relocate apparently at random in different environments. This phenomenon is commonly observed in rodent hippocampal cells and is termed *global remapping* (38; 53; 54).

After considering random trajectories in space, we then consider behaviour more like animal i.e. spending time near boundaries and approaching objects. Here, because the transition statistics change,

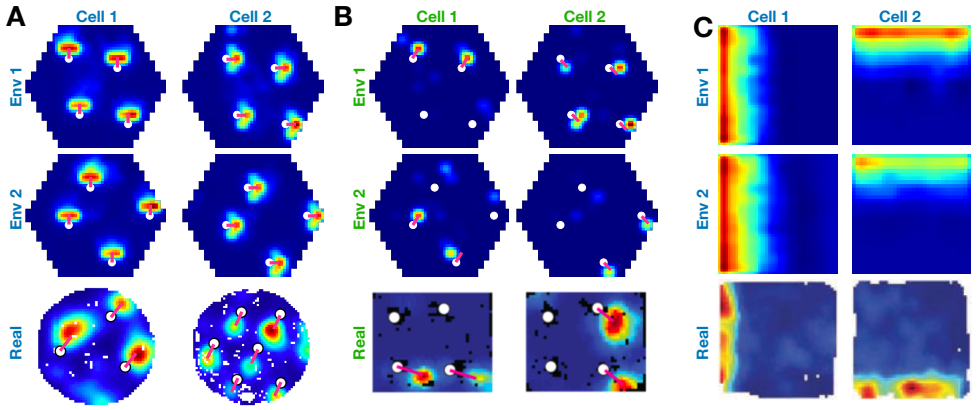

Figure 5: **TEM learned representations reflect transition statistics.** When the agent's transition statistics mimic different behaviours, TEM learns new representations (left to right: different cells, top to bottom: environments 1, 2, real data). **A)** When biased to move towards objects (white dots) TEM learns structural cells with a vector relationship to the objects - object vector cells (26). These cells generalise to *all* objects. **B)** TEM hippocampal cells reflect this behavioural transition change with similar cells, though they do not generalise to all objects - landmark cells (24). **C)** When biased towards boundaries, TEM learns border cell like representations (27).

so do the optimal representations for predicting future location. Now medial entorhinal representations, $\mathbf{g}$, in TEM include border cells (27) (Figure 5C) and cells that fire at the same distance and angle from any object (object vector cells (26); Figure 5A) for the two cases respectively.

Similar cells exist in TEM's hippocampal layer, $\mathbf{p}$, with a crucial difference. Here, object sensitive cells represent the vector to particular objects but do not generalise across objects (Figure 5B) - they represent the conjunction between the structural representation and the sensory data. These cells are reminiscent of 'landmark' cells that have been recorded in rodent hippocampus (24).

In the remainder of this chapter we consider neural representations in tasks that have both spatial and non-spatial components. We consider a recent finding by Sun et al. (66). Rodents perform laps of a circular track, but only receive reward every four laps. Now hippocampal cells develop a new representation. Whilst some cells represent location on the track, (i.e. place cells; Figure 6A top), others are also spatially selective, but fire only on one of the 4 laps (Figure 6A middle). A third set fire again at a set spatial location, but vary their firing continuously as a function of lap number (Figure 6A bottom). Hippocampal cells maintain a complex combined representation of space and lap number.

When TEM was trained on this task, it learned these same 3 representations in hippocampus (Figure 6B). Importantly, TEM allows us to reveal a candidate mechanism. TEM's medial entorhinal cells have reconfigured to code differently for each lap, understanding that the abstract task space is not a single lap but four (Figure 6C bottom). These results suggest entorhinal cells can learn to represent tasks at multiple levels of cognitive abstraction simultaneously, with hippocampal cells reflecting their conjunction with sensory experience.

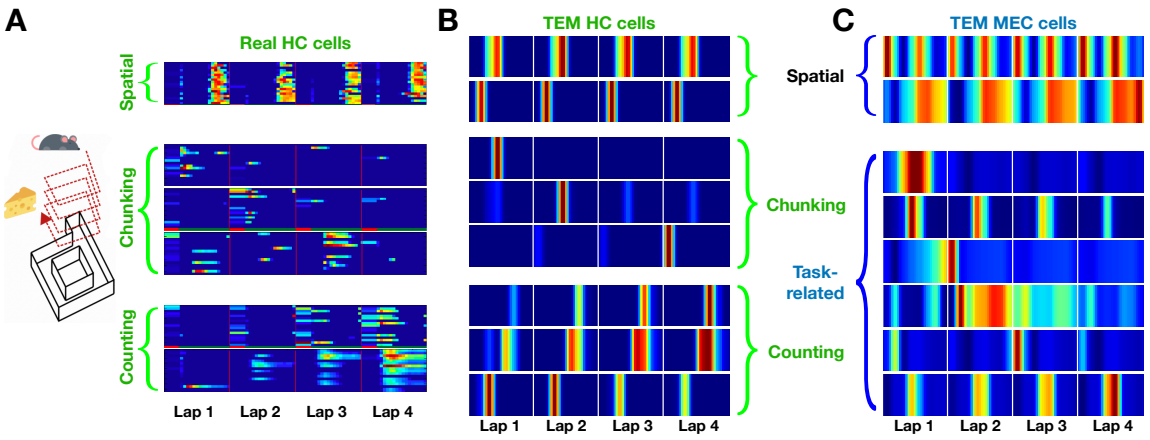

Figure 6: **TEM represents non-spatial reinforcement learning tasks and predicts non-spatial remapping. A)** In (66), rodents perform laps of a track, only 'rewarded' every 4 laps. Different hippocampal cell types are found: spatial place-like cells (top), those that preferentially fire on a given lap (middle) and those that count laps (bottom). **B)** TEM learns similar representations when only 'rewarded' every 4 laps. **C)** TEM medial entorhinal cells learn both spatially periodic cells (top), and cells that represent the non-spatial task structure of 'every 4 laps' (bottom). The latter cells are yet experimentally observed, but are predicted by TEM.

Chapter 8: The Tolman-Eichenbaum Machine for machine learners

Chapter 8 describes TEM in more detail. Specifically a complete mathematical description of the model is provided, which we do not delve into further here.

Chapter 9: Neural predictions from TEM

Chapter 9 looks at a neural prediction of TEM. Contrary to popular belief, TEM proposes that structural knowledge between grid and place cells is preserved, rather than random remapping of the place code. This prediction is tested and verified in simultaneous recordings of place and grid cells in a remapping experiment.

We tested data from two experiments in which both place and grid cells have been recorded whilst rats (67) and mice (68) freely forage in multiple environments. We demonstrated that the activity of each grid cell at the peak firing of each place cell peak (gridAtPlace) is correlated across environments in both datasets (Figure 7.

These results demonstrate non-random place cell remapping in space, and support a key prediction of our model: that hippocampal place cells, despite remapping across environments, retain their relationship with the entorhinal grid, providing a substrate for structural inference.

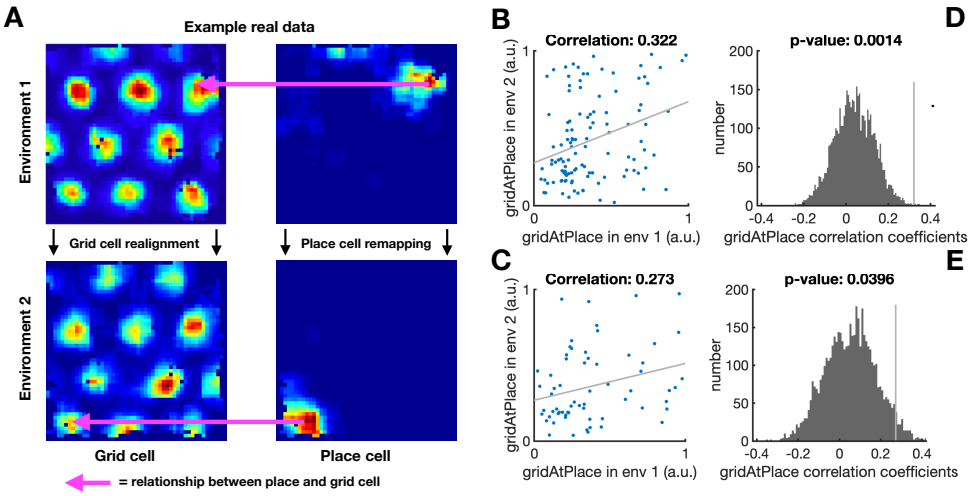

Figure 7: **Structural knowledge is preserved over apparently random hippocampal remapping. A)** TEM predicts place cells remap to locations consistent with a grid code, i.e. a place cell co-active with a grid cell will be more likely to remap to locations where that grid cell is also active. **B-C)** Data from open field remapping experiments with simultaneously recorded place and grid cells (67; 68). We compute the grid cell firing rate at the location of place cell peak for every grid cell - place cell pair in each of the two environments, and then correlate this measure across environments (left). We compare this correlation coefficient to those computed equivalently but with randomly permuted place cell peaks (right). This is done for two independent datasets **B)** (67) and **C)** (68). The true observed correlation coefficients lies off the null distribution ($p < 0.05$), demonstrating place cell remapping is not random, but instead tied to structural constraints of grid cells.

Chapter 10: Afterword

Building an understanding that spans from computation through cellular activity to behaviour is a central goal of neuroscience. This thesis attempts to bridge these levels of understanding by translating high level ideas underpinned with Bayesian reasoning into neural algorithms and representations. The first part showed that framing supervised learning probabilistically allows for approximate inference to be implemented in biological neural networks. Furthermore, learning in these networks is equivalent to the famous back-propagation algorithm. This work demonstrated that algorithms as efficient as those in machine learning could be implemented in the brain. The second part presented a framework for learning and generalising abstract relational knowledge. This framework unifies spatial cognition with relational memory in the hippocampal formation. A model implementation, TEM, learned to represent relational knowledge using neural representations that mirror those found in the brain. Furthermore, a prediction of TEM was verified in neural recordings.

We finish with a prospective outlook on future questions for learning algorithms and generalising representations in the brain. These challenges present the neuroscience and machine learning community with difficult, yet interesting times ahead. Solving them would offer deep insights into the nature of intelligence, the human mind, and ultimately consciousness.

REFERENCES

C. S. Sherrington, "The Integrative Action of the Nervous System.," *JAMA: The Journal of the American Medical Association*, vol. XLVIII, p. 1055, mar 1907.

O. Loewi, "Über humorale übertragbarkeit der Herznervenwirkung," *Pflügers Archiv für die Gesamte Physiologie des Menschen und der Tiere*, vol. 189, pp. 239–242, dec 1921.

W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, dec 1943.

A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, pp. 500–544, aug 1952.

E. L. Thorndike, "Animal intelligence: An experimental study of the associative processes in animals.," *The Psychological Review: Monograph Supplements*, vol. 2, no. 4, pp. i–109, 1898.

I. P. Pavlov, "Conditioned reflexes," 1927.

B. F. Skinner, *The behavior of organisms*. Appleton-Century, 1938.

C. B. Ferster and B. F. Skinner, *Schedules of reinforcement.* East Norwalk: Appleton-Century-Crofts, 1957.

J. J. Gibson, "The theory of affordances," 1977.

D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, "A General Framework for Parallel Distributed Processing," *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume I*, no. 1982, pp. 45–76, 1986.

P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. 1994.

J. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*. jun 1978.

T. Hafting, M. Fyhn, S. Molden, M.-b. B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.

I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a Definition of Disentangled Representations," pp. 1–29, dec 2018.

Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, B. Marc G, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, G. Wayne, H. Soyer, F. Viola, B. Zhang, R. Goroshin, N. Rabinowitz, R. Pascanu, C. Beattie, S. Petersen, A. Sadik, S. Gaffney, H. King, K. Kavukcuoglu, D. Hassabis, R. Hadsell, and D. Kumaran, "Vector-based navigation using grid-like representations in artificial agents," *Nature*, vol. 557, pp. 429–433, may 2018.

J. C. R. Whittington, T. H. Muller, S. Mark, C. Barry, and T. E. J. Behrens, "Generalisation of structural knowledge in the hippocampal-entorhinal system," *Advances in Neural Information Processing Systems 31*, vol. 31, pp. 8493–8504, 2018.

D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, 2016.

J. S. Bowers, "Parallel Distributed Processing Theory in the Age of Deep Networks," *Trends in Cognitive Sciences*, vol. 21, no. 12, pp. 950–961, 2017.

J. C. R. Whittington and R. Bogacz, "An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity," *Neural Computation*, vol. 29, pp. 1229–1262, may 2017.

J. C. R. Whittington and R. Bogacz, "Theories of Error Back-Propagation in the Brain," *Trends in Cognitive Sciences*, vol. xx, pp. 1–16, 2019.

S. S. Deshmukh and J. J. Knierim, "Influence of local objects on hippocampal representations: Landmark vectors and memory.," *Hippocampus*, vol. 23, pp. 253–67, apr 2013.

J. Krupic, N. Burgess, and J. O'Keefe, "Neural Representations of Location Composed of Spatially Periodic Bands," *Science*, vol. 337, pp. 853–857, aug 2012.

Ø. A. Høydal, E. R. Skytøen, S. O. Andersson, M.-B. Moser, and E. I. Moser, "Object-vector coding in the medial entorhinal cortex," *Nature*, vol. 568, pp. 400–404, apr 2019.

T. Solstad, C. N. Boccara, E. Kropff, M.-B. Moser, and E. I. Moser, "Representation of Geometric Borders in the Entorhinal Cortex," *Science*, vol. 322, pp. 1865–1868, dec 2008.

B. A. Strange, M. P. Witter, E. S. Lein, and E. I. Moser, "Functional organization of the hippocampal longitudinal axis," *Nature Reviews Neuroscience*, vol. 15, no. 10, pp. 655–669, 2014.

T. Danjo, T. Toyoizumi, and S. Fujisawa, "Spatial representations of self and other in the hippocampus," *Science*, vol. 359, no. 6372, pp. 213–218, 2018.

D. B. Omer, S. R. Maimon, L. Las, and N. Ulanovsky, "Social place-cells in the bat hippocampus," *Science*, vol. 359, no. 6372, pp. 218–224, 2018.

J. Taube, R. Muller, and J. Ranck, "Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis," *The Journal of Neuroscience*, vol. 10, pp. 420–435, feb 1990.

Ø. A. Høydal, E. R. Skytøen, M. B. Moser, and E. I. Moser, "Object-vector cells in the medial entorhinal cortex," *bioRxiv*, vol. 0, 2018.

J. L. Gauthier and D. W. Tank, "A Dedicated Population for Reward Coding in the Hippocampus," *Neuron*, vol. 99, no. 1, pp. 179–193.e7, 2018.

C. Lever, S. Burton, A. Jeewajee, J. O'Keefe, and N. Burgess, "Boundary vector cells in the subiculum of the hippocampal formation," *Journal of Neuroscience*, vol. 29, no. 31, pp. 9771–9777, 2009.

A. Sarel, A. Finkelstein, L. Las, and N. Ulanovsky, "Vectorial representation of spatial goals in the hippocampus of bats," *Science*, vol. 355, no. 6321, pp. 176–180, 2017.

T. E. J. Behrens, T. H. Muller, J. C. R. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-nelson, "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior," *Neuron*, vol. 100, no. 2, pp. 490–509, 2018.

M. Fyhn, T. Hafting, A. Treves, M. B. Moser, and E. I. Moser, "Hippocampal remapping and grid realignment in entorhinal cortex," *Nature*, vol. 446, no. 7132, pp. 190–194, 2007.

E. Bostock, R. U. Muller, and J. L. Kubie, "Experience-dependent modifications of hippocampal place cell firing.," *Hippocampus*, vol. 1, no. 2, pp. 193–205, 1991.

S. Leutgeb, J. K. Leutgeb, C. A. Barnes, E. I. Moser, B. L. McNaughton, and M.-B. Moser, "Independent Codes for Spatial and Episodic Memory in Hippocampal Neuronal Ensembles," pp. 619–624, jul 2005.

A. O. Constantinescu, J. X. OReilly, T. E. J. Behrens, J. X. O'Reilly, and T. E. J. Behrens, "Organizing conceptual knowledge in humans with a gridlike code," *Science*, vol. 352, pp. 1464–1468, jun 2016.

J. B. Julian, A. T. Keinath, G. Frazzetta, and R. A. Epstein, "Human entorhinal cortex represents visual space using a boundary-anchored grid," *Nature Neuroscience*, vol. 21, no. 2, pp. 191–194, 2018.

M. Nau, T. Navarro Schröder, J. L. S. Bellmund, and C. F. Doeller, "Hexadirectional coding of visual space in human entorhinal cortex," *Nature Neuroscience*, vol. 21, no. 188, 2018.

J. A. Dusek and H. Eichenbaum, "The hippocampus and memory for orderly stimulus relations," *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 7109–7114, 1997.

D. Kumaran, H. L. Melo, and E. Duzel, "The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies," *Neuron*, vol. 76, no. 3, pp. 653–666, 2012.

J. C. R. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. E. Behrens, "The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation," *Cell*, vol. 183, pp. 1249–1263.e23, nov 2020.

C. Kemp and J. B. Tenenbaum, "The discovery of structural form," *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10687–10692, 2008.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," *International Conference on Learning Representations*, vol. 0, jul 2017.

J. R. Manns and H. Eichenbaum, "Evolution of declarative memory.," *Hippocampus*, vol. 16, no. 9, pp. 795–808, 2006.

C. J. Cueva and X.-X. Wei, "Emergence of grid-like representations by training recurrent neural networks to perform spatial localization," *International Conference on Learning Representations*, vol. 0, mar 2018.

K. Yoon, M. A. Buice, C. Barry, R. Hayman, N. Burgess, and I. R. Fiete, "Specific evidence of low-dimensional continuous attractor dynamics in grid cells," *Nature Neuroscience*, vol. 16, no. 8, pp. 1077–1084, 2013.

M. W. Jung, S. I. Wiener, and B. L. McNaughton, "Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat," *The Journal of Neuroscience*, vol. 14, no. 12, pp. 7347–7356, 1994.

K. B. Kjelstrup, T. Solstad, V. H. Brun, T. Hafting, S. Leutgeb, M. P. Witter, E. I. Moser, and M.-B. Moser, "Finite Scale of Spatial Represenation in the Hippocampus," *Science*, vol. 321, no. July, pp. 140 – 143, 2008.

M. I. Anderson and K. J. Jeffery, "Heterogeneous modulation of place cell firing by changes in context," *Journal of Neuroscience*, vol. 23, no. 26, pp. 8827–8835, 2003.

R. U. Muller and J. L. Kubie, "The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells," *Journal of Neuroscience*, vol. 7, no. 7, pp. 1951–1968, 1987.

E. M. Purcell, "Life at low Reynolds number," *American Journal of Physics*, vol. 45, no. 1, pp. 3–11, 1977.

W. B. Scoville and B. Milner, "Loss of recent memory after bilateral hippocampal lesions," *J.Neurol. Neurosurg.Psychiat.*, vol. 20, pp. 11–21, 1957.

H. Eichenbaum and N. J. Cohen, "Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function?," *Neuron*, vol. 83, no. 4, pp. 764–770, 2014.

H. Eichenbaum, P. Dudchenko, E. Wood, M. Shapiro, and H. Tanila, "The Hippocampus, Memory, Review and Place Cells: Is It Spatial Memory or a Memory Space? might occur at different locations. Olton and colleagues Neuron 210 Figure 1. Schematic Overhead Views of Four Different Types of Apparatus and Examples of Location-S," *Neuron*, vol. 23, pp. 209–226, 1999.

C. Lever, T. Wills, F. Cacucci, N. Burgess, and J. O'Keefe, "Long-term plasticity in hippocampal place-cell representation of environmental geometry," *Nature*, vol. 416, pp. 90–94, mar 2002.

M. Bunsey and H. Eichenbaum, "Conservation of hippocampal memory function in rats and humans," *Nature*, vol. 379, pp. 255–257, jan 1996.

J. L. McClelland, B. L. McNaughton, and R. 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," *Psychological Review*, vol. 102, no. 3, pp. 419–457, 1995.

I. Momennejad, "Learning Structures: Predictive Representations, Replay, and Generalization," *Current Opinion in Behavioral Sciences*, vol. 32, pp. 155–166, apr 2020.

K. L. K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman, "The hippocampus as a predictive map," *Nature Neuroscience*, vol. 20, no. 11, pp. 1643–1653, 2017.

E. C. Tolman, "Cognitive maps in rats and men.," *Psychological Review*, vol. 55, no. 4, pp. 189–208, 1948.

H. Stensola, T. Stensola, T. Solstad, K. FrØland, M.-B. B. Moser, and E. I. Moser, "The entorhinal grid map is discretized," *Nature*, vol. 492, no. 7427, pp. 72–78, 2012.

C. Sun, W. Yang, J. Martin, and S. Tonegawa, "Hippocampal neurons represent events as transferable units of experience," *Nature Neuroscience*, vol. 23, pp. 651–663, may 2020.

C. Barry, L. L. Ginzberg, J. O'Keefe, and N. Burgess, "Grid cell firing patterns signal environmental novelty by expansion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 43, pp. 17687–17692, 2012.

G. Chen, J. A. King, Y. Lu, F. Cacucci, and N. Burgess, "Spatial cell firing during virtual navigation of open arenas by head-restrained mice," *eLife*, vol. 7, 2018.