# Neural network architectures for relational reasoning, planning, and cognitive intelligence (D3.12 – SGA3)

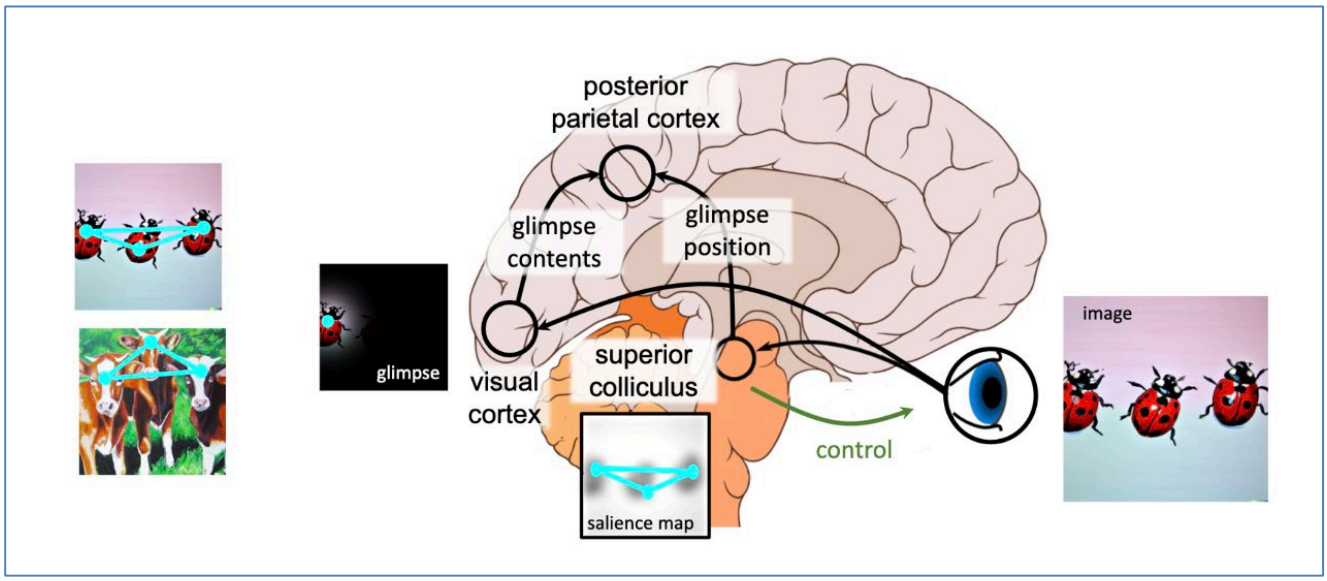


**Figure 1: Enactive theory of relational reasoning applied to counting**

The image shows the brain network and theory that inspired our deep neural network model of zero-shot counting.

| Project Number: | 945539 | Project Title: | HBP SGA3 |
|---|---|---|---|

| | |
|---|---|
| Document Title: | Neural network architectures for relational reasoning, planning, and cognitive intelligence |
| Document Filename: | D3.12 (D31) SGA3 M42 RESUBMITTED 231208.docx |
| Deliverable Number: | SGA3 D3.12 (D312) |
| Deliverable Type: | Other |
| Dissemination Level: | PU = Public |
| Planned Delivery Date: | SGA3 M42 / 30 SEP 2023 |
| Actual Delivery Date: | SGA3 M42/ 27 SEP 2023 (resubmitted 08 Dec 2023) |
| Author(s): | Christopher SUMMERFIELD, UOXF (P59) |
| Compiled by: | Christopher SUMMERFIELD, UOXF (P59) |
| Contributor(s): | |
| WP QC Review: | Jeannette BOSCHMA, UM (P117); Anita VAN OERS, UM (P117) |
| WP Leader / Deputy Leader Sign Off: | Rainer GOEBEL, UM (P117) |
| T7.4 QC Review: | N/A |
| Description in GA: | A neural network architecture will be built based on the neurobiology of the primate prefrontal cortex. Its performance will be evaluated on verbal and/or nonverbal intelligence tests. The deliverable will consist of (i) an evaluation dataset; (ii) the network code; and providing (iii) a summary of results achieved. |
| Abstract: | D3.12 aimed to build models of complex processes, such as reasoning, planning and compositional inference, by drawing on our understanding of computational principles in biological neural networks. We trained neural network models to solve tasks that involved counting the number of unique elements in a scene, planning to get to a goal in a multi-compartment environment, and combining two pieces of abstract knowledge to make an inference. |
| Keywords: | Neural networks, cognition, relational reasoning, planning, navigation, composition |
| Target Users/Readers: | Computational Neuroscientists, Machine Learning researchers, Cognitive scientists, interested public |

## Table of Contents



## Table of Figures



## History of Changes made to this Deliverable (post Submission)

| Date | Change Requested / Change Made / Other Action |
|---|---|
| 27 SEP 2023 | Deliverable submitted to EC |
| | Resubmission with specified changes requested in Review Report<br>Main changes requested:<br>The deliverable is acceptable in terms of scientific content, but the quality of the text is not sufficient and needs to be revised before public release. Either add details from or clearly list links to peer-reviewed publications or preprints related to the reported outcomes. |
| | Revised draft sent by WP to PCO.<br>Main changes made, with indication where each change was made:<br>• added hyperlinks to all the references in the bibliography |
| | Revised version resubmitted to EC by PCO via SyGMa |

# 1. Introduction

The focus of D3.12 has been to build neural networks that display cognitive abilities including reasoning, planning, and compositional inference. Recent years have seen remarkable progress in AI research. The advent of large generative models has thrust issues of artificial reasoning and compositional inference into the limelight. However, current AI systems lack core functionality that is present in the human brain. In particular, they are excessively data-hungry, lack robustness, and use computational motifs that may be very different from those in biology. In In task T3.6, the goal has been to explore how neural network models can recreate the systematic behaviour of humans displayed by humans when reasoning or planning.

Over recent years, there has been a blossoming of interest in neuroscience for tools, concepts and datasets from machine learning. In particular, we have witnessed a renaissance of interest in deep neural network models of vision and audition, which have used off-the shelf architectures as models of perception. In task T3.6, we take a different approach, by conducting systematic experiments to understand how neural networks might be capable of systematic behaviours. This contributes to HBP focus on Cognitive Functions and Neural Networks.

We have focussed on three key outstanding problems in AI research, using ideas inspired by biology, and specifically human neuroscience. The questions were as follows:

- How do intelligent agents understand visual scenes? [OP3.15]
- How do intelligent agents plan during navigation? [OP3.16]
- How do intelligent agents efficiently encode new knowledge? [OP3.17]

Over the course of SGA3, UOXF has tackled these challenges in three parallel streams. These have been conducted collaboration with TUGRAZ (P55) (Wolfgang Maas), TUM (Fabrice Morin) and UM (P117) (Mario Senden).

The work will be of interest to a broad community of researchers spanning cognitive science, computational neuroscience, machine learning, and AI research.

# 2. Relational Reasoning

Visual scenes are understood not just by the objects they contain, but by the relations between these objects. For example, in Figure 2a, the four panels each contain two objects: a man and a car. However, grasping the relationship between the man and the car is vital for understanding what is occurring in the scene. Current AI systems, based on deep convolutional neural networks (CNNs), excel at the recognition of lone objects with a single, rapid glimpse. However, systems currently struggle to interpret scenes by processing the relations between objects. In P2200 (Summerfield et al., 2020), we developed a theory of the shortcomings of current AI systems and how biology tackles this problem. In OP3.15, we focussed on a simple relational property of visual scenes – numerosity. We chose numerosity because it is a domain where AI systems – including state-of-the-art large generative models – still perform poorly. The specific challenge that we set out to solve is called “zero-shot counting” (See Figure 1). We asked how it might be possible to train a neural network to accurately count objects drawn from a class that it had never seen before, a challenge that human can effortlessly solve.

To address this challenge, we drew inspiration from the dual-streams architecture of the human brain, as well as enactive theories of learning from cognitive psychology. We built a neural network that viewed a visual scene through a series of discrete glimpses, mirroring the saccadic system in the primate brain (Figure 2b). Like the primate brain, the network receives information about both what (the glimpse contents) and where (the glimpse location), which allows it to form representations that multiplex information about object identity and relational spatial information during representation learning. We predicted that this model would be able to solve zero-shot counting problems, and that in doing so, it would recreate key aspects of the neurophysiology of the primate number system.

The project was successful, and has now been described in conference proceedings P4055 and P4142 (Thompson et al., 2023, 2022), and is in the late stages of being written up for a journal article. The results are summarised below. Unlike state-of-the-art machine learning models of the human visual systems, like CNNs, the model is able to solve zero-shot counting problems, but model lesion studies reveal that this depends on the integrity of both dorsal and ventral streams. In solving this problem, the network forms neural representations of space and number that are very similar to those observed in primate area LIP; its pattern of representation learning mirrors well-described developmental trajectories known from cognitive psychology, and it predicts counting performance with and without distraction in human participants (Figure 2c-g).

Data and deep neural network models for relational reasoning can be found here:

https://github.com/summerfieldlab/saccades.

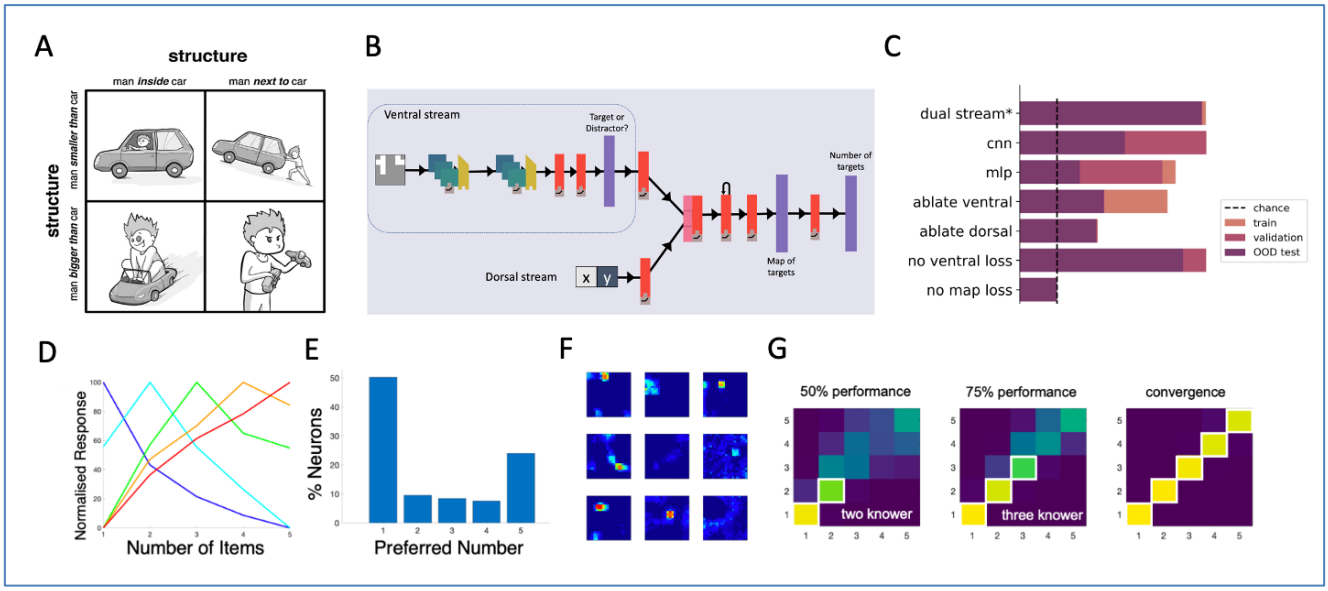


**Figure 2: Deep neural networks for relational reasoning**

**(a) Relational information is important for scene understanding. (b) Schematic of dual streams network. (c) Performance of dual streams network on OOO tests (purple). (d-f) Lognormal codes for number in the RNN, distributions of preferred numerosities, and spatial receptive fields in the network match those in primate LIP. (g) Developmental trajectories are similar to those in children, with lower numbers learned first.**

# 3. Planning in spatial navigation

Current AI systems continue to display limited versatility, especially in spatiotemporally extended environments such as video games. One key outstanding challenge is the question of how intelligent agents plan a route to a goal. During navigation, agents including humans and rodents encode the environment with diverse classes of spatially sensitive neuron housed in the hippocampal-entorhinal cortex, including place and grid cells. However, much less is known about how goals are encoded in the human brain, and how this coding scheme may permit versatile, context-dependent navigation. Here, we tackled this question using a classic AI problem implemented as a video game (Figure 3a-b).

In OP3.16, we developed a new model of context-dependent navigation, based on the functional properties of the human medial temporal lobe. It proposes that an agent navigates using a combination of place cells (encoding its current location) and goal cells (encoding the location of its current destination, which can be used for spatial planning; Figure 3c). Critically, the goal cells can change flexibly with context, allowing the agent to navigate to distinct destinations based on a contextual cue. The model makes a striking prediction about the neural geometry of the medial temporal lobe representation: that neural signals should be compressed along spatial dimensions that span current goals (Figure 3d). We tested this by asking human participants to carry out a variant of the classic goal-directed navigation task used in AI research, called the Four Rooms Task, which

involve navigating to two goals in succession. Humans solving this task formed exactly the sorts of representations predicted by the goal-directed model. This work is now published (P4137) (Muhle-Karbe et al., 2023).

In a complementary set of experiments, the team at UGRAZ have studied planning by developing a biologically inspired neural network that can solve planning problems in high-dimensional graph-like environments. Unlike deep networks, it does so using only local prediction rules, and a theoretical framework that is closely related to predictive coding. The network has been translated to solve both navigation problems (Figure 3e), and the control of locomotion in simulated quadruped (Figure 3f). This work is described in P4037 (Stöckl and Maass, 2022).

Models for spatial planning can be found here:

https://zenodo.org/record/8246406

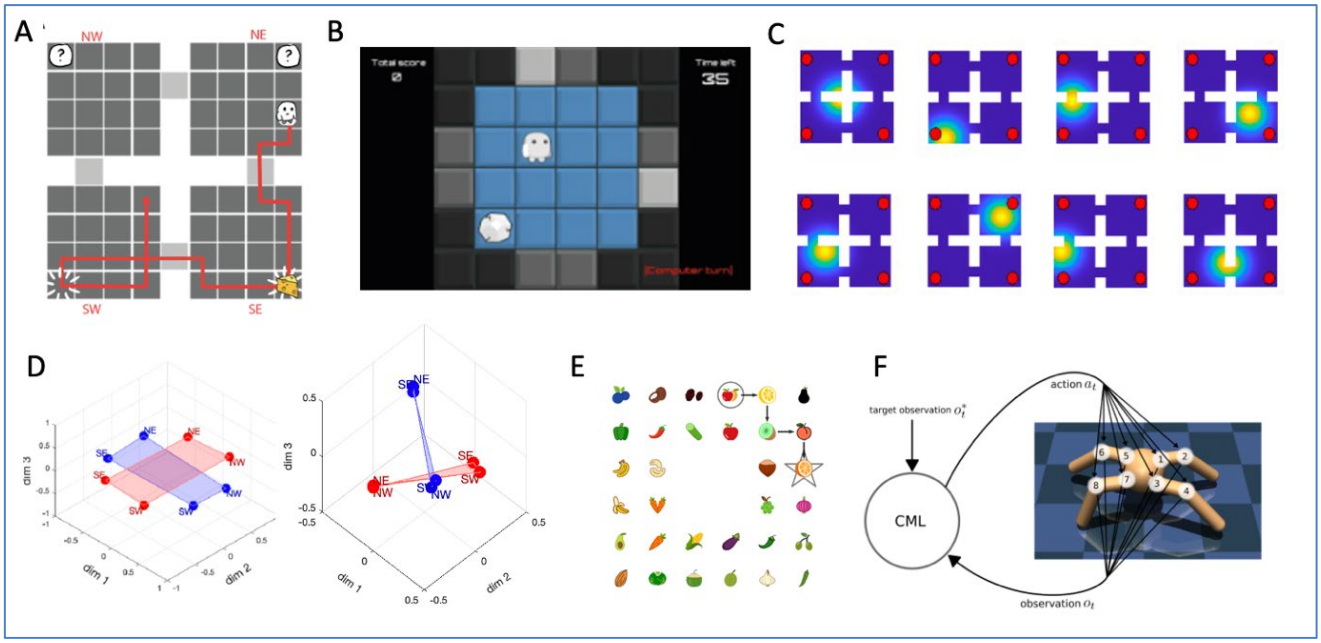


**Figure 3: Models of spatial planning**

**(a) The four rooms environment, with two successive goals. (b) Screen shot of the video game. (c) Example place and goal cells in the model. (d) Predicted (left) and observed (right) neural geometries in hippocampus. (e-f) Problems solved by the high-dimensional planner.**

# 4. Cognitive Intelligence

A core challenge for current AI systems is how to efficiently encode knowledge in ways that support intelligent behaviour. For example, AI systems remain susceptible to catastrophic interference, whereby new knowledge acquisition interferes with existing performance. Solving this *continual learning* problem remains one of the key outstanding challenges in AI research. In OP3.17, we tackled this problem via a deceptively simple reasoning task on which current neural networks fail dramatically. The task requires agents to first learn two transitive series (e.g. to learn that A1>B1>C1 and A2>B2>C2) and then, from just a handful of trials (few-shot learning) in which they are taught that C1>A2, to infer the relations between all items in the set. Standard feedforward neural networks fail dramatically at this task; they learn that C1>A2 is an exception, and do not generalise this knowledge to the reorganise their full knowledge structure. Monkeys are known to be able to solve this task, but only after extensive practice.

In OP3.17, we studied how humans performed the task and used fMRI to study how their representations adjusted after few-shot learning. We found that humans readily reorganised their knowledge structures after just a few shots, and that neural representations in the parietal cortex changed accordingly. This builds on our previous work showing that the parietal cortex is a key region

for relational inference over low-dimensional data structures (P3070) (Luyckx et al., 2019) as well as other work on continual learning (P4139 and P4140) (Flesch et al., 2022a, 2022b). Building on these human data, and on more general principles of biological learning, we applied a new computational principle to standard neural networks, which involves a stabilisation mechanism: frequently encountered associations are tagged as being stable, so that during few-shot learning, updates are applied not only to the experienced items (C1 and A2) but also to their relevant associates (Figure 4a). When the stabilisation parameter was low, the networks reduced to standard feedforward networks; when it was high, the networks solved the inference problem like the best human performers; and when it was intermediate, they performed like the more poorly performing humans (Figure 4b). This work, which is described in P4138 (Nelli et al., 2023), thus describes a new biologically inspired mechanism for solving complex inference problems.

Data and deep neural network models for cognitive intelligence can be found here:

https://github.com/mwhitemfldm/CompositionalNumericalReasoning/tree/main

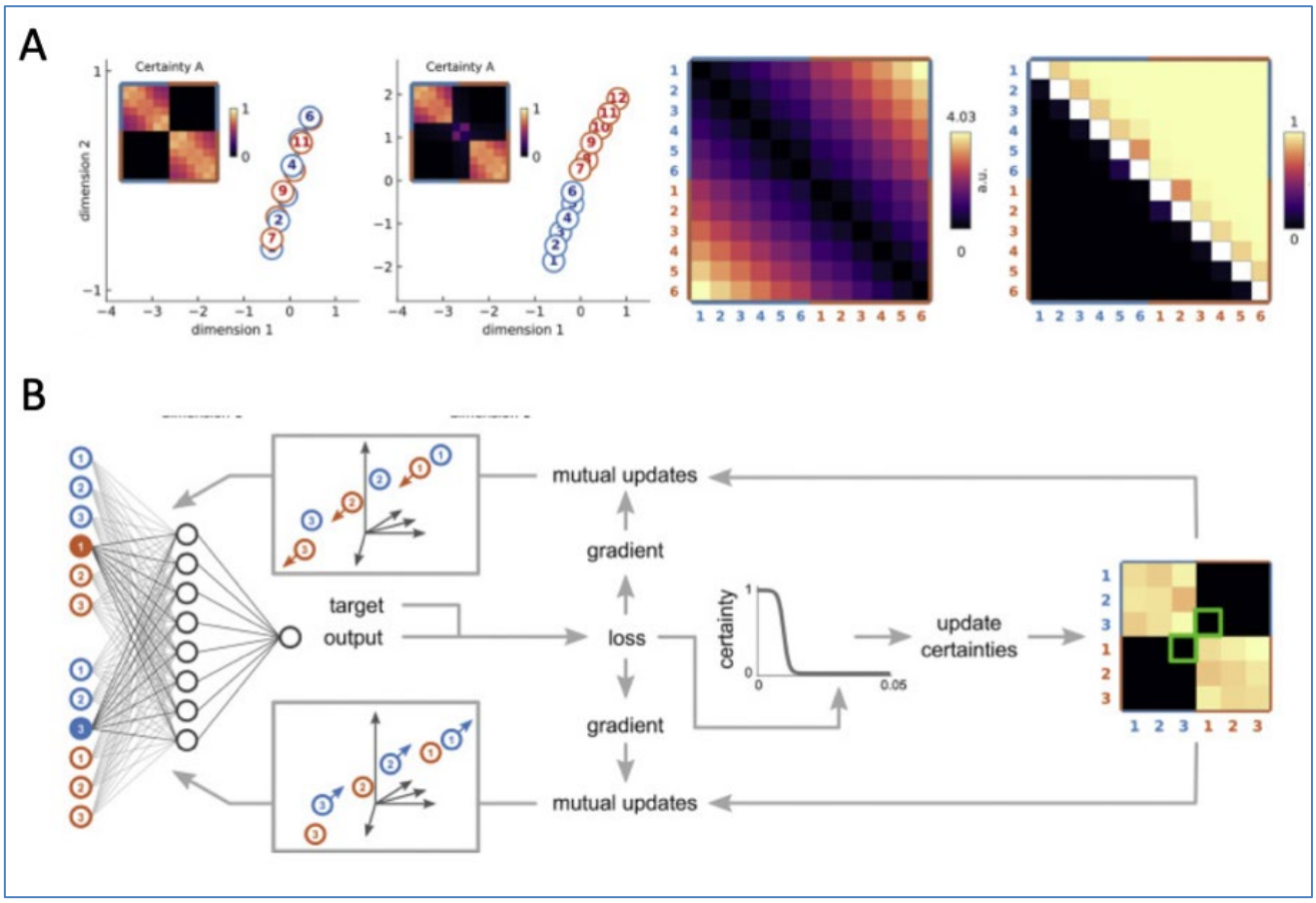


**Figure 4: A neural network for knowledge assembly**

**(a) Examples of the internal representations of stability in the model, and model transitive inference performance. (b) Architecture and approach in our model.**

## 4.1 Generalised Latent Equilibrium

In D3.10 and D3.11 UBERN (P71) presented Latent Equilibrium (LE), a novel framework for computation and learning in biological and artificial neural networks. LE is particularly suited to process static input such as images of handwritten digits (e.g., the MNIST dataset). In such scenarios, the coding is purely spatial, so the time dimension only plays a role when input patterns are switched.

Our daily environment, however, is inherently dynamic and sensory inputs are constantly changing. In such an environment, information processing becomes even more difficult. For example, in order

to recognise a sound (or, even more difficultly, a song), it is not only important to know which notes are played, but also at which time they are played and for how long. In order to learn in such inherently dynamic environments, our brains need to solve the spatio-temporal credit assignment problem: "How to adapt the right synapses given their input at multiple points in time in order to solve a given problem?"

In Machine Learning (ML), there exist well-known solutions that enable learning from time-continuous inputs, such as Backpropagation Through Time (BPTT) or Real-Time Recurrent Learning (RTRL). While these algorithms allow to learn from spatio-temporal data with Artificial Neural Networks (ANNs), they have multiple shortcomings; either using learning rules that are non-local (in time) or that require too much memory, which is why they do not scale well to large networks. Partial solutions exist, such as e-prop (Bellec et al., 2019 – P1998), but they rely on parametrizations that may not be compatible with biological observations, while still maintaining a relatively high resource cost.

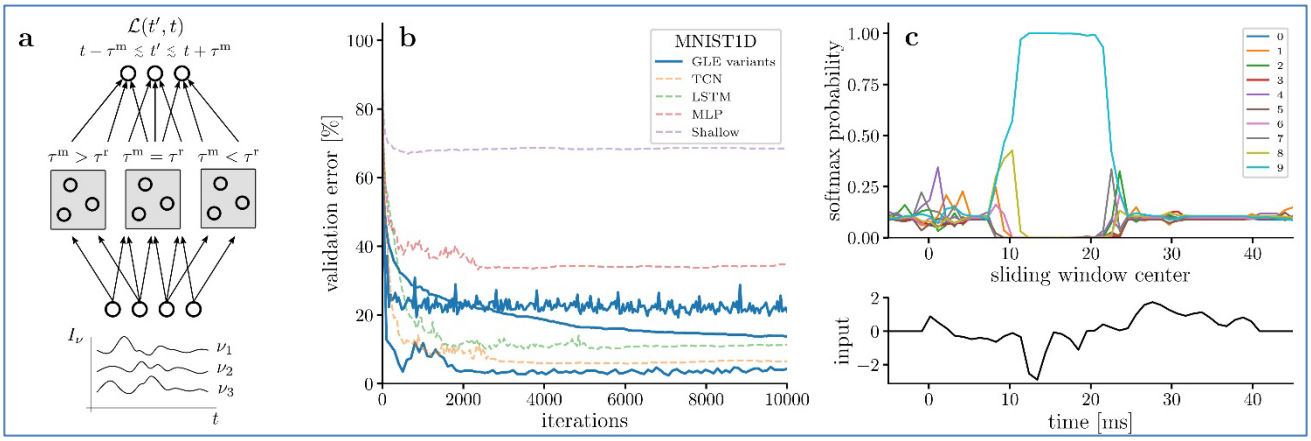


**Figure 5: Preliminary results of the GLE framework.**

**(a) Sketch of a network able to process time-continuous signals within the GLE framework. (b) Performance of different GLE variants on the MNIST1D dataset (cf. Greydanus, 2020). Different GLE variants differ in whether they use lags or true delays as well as whether their architecture is partly hand-engineered or if they are trained end-to-end. (c) Exemplary SoftMax output of the network during classification of the digit 9 using an ANN. The input is streamed in such a way that the instantaneous network sees a temporal window of 10 ms at every moment.**

To overcome these difficulties, we are currently working on a generalisation of our Latent Equilibrium framework towards combining quasi-instantaneous computation with the transient memory of slow membrane voltages to enable information processing of spatio-temporal inputs (see Figure 5). Incorporating memory within our framework allows the mapping of temporal problems (which so far could only be tackled via BPTT or approximations thereof) to spatial problems, which can then be learned with appropriate modifications of LE. In preliminary simulations, we tested the capabilities of networks of neurons with multiple time constants to learn temporal patterns on multiple time scales and early results look very promising. We therefore envision this generalised framework to provide a comprehensive solution for spatio-temporal credit assignment in dynamical physical systems and we will carry on this research beyond the HBP.

# 5. Looking Forward

There remains much work to be done. Since its inception, the HBP has overlooked Natural Language Processing (NLP), the study of how to build artificial systems capable of the quintessential human cognitive capacity, the ability to speak in structured sentences. The success of large-scale transformer-based generative models has opened the door to new lines of research in which reasoning, planning and composition are studied in the domain of natural language. An important avenue for future research will be to scale the work described here, to ask whether the insights are relevant for new AI technologies that operate at scale.

# 6. References

Bellec, G., Scherr, F., Subramoney, A. et al. A solution to the learning dilemma for recurrent networks of spiking neurons. Nat Commun 11, 3625 (2020). https://doi.org/10.1038/s41467-020-17236-y (P1998)

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., Summerfield, C., 2022a. Orthogonal representations for robust context-dependent task performance in brains and neural networks. Neuron S0896627322000058. https://doi.org/10.1016/j.neuron.2022.01.005 (P413

Flesch, T., Nagy, D.G., Saxe, A., Summerfield, C., 2022b. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. https://doi.org/10.48550/ARXIV.2203.11560 (P4140)

Luyckx, F., Nili, H., Spitzer, B., Summerfield, C., 2019. Neural structure mapping in human probabilistic reward learning. Elife 8. https://doi.org/10.7554/eLife.42816 (P3070)

Muhle-Karbe, P.S., Sheahan, H., Pezzulo, G., Spiers, H.J., Chien, S., Schuck, N.W., Summerfield, C., 2023. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. Neuron S0896627323006323. https://doi.org/10.1016/

Nelli, S., Braun, L., Dumbalska, T., Saxe, A., Summerfield, C., 2023. Neural knowledge assembly in humans and neural networks. Neuron S0896627323001186. https://doi.org/10.1016/j.neuron.2023.02.014 (P4138)

Stöckl, C., Maass, W., 2022. Local prediction-learning in high-dimensional spaces enables neural networks to plan (preprint). Neuroscience. https://doi.org/10.1101/2022.10.17.512572 (P4037)

Summerfield, C., Luyckx, F., Sheahan, H., 2020. Structure learning and the posterior parietal cortex. Progress in neurobiology 184, 101717. 10.1016/j.pneurobio.2019.101717 (P2200)

Thompson, J.A.F., Sheahan, H., Summerfield, C., 2023. Zero-Shot Visual Numerical Reasoning in Dual-Stream Neural Networks, in: Proceedings of the Cognitive Computational Neuroscience Society 2023. Presented at the Cognitive Computational Neuroscience, Oxf

Thompson, J.A.F., Sheahan, H., Summerfield, C., 2022. Learning to count visual objects by combining “what” and “where” in recurrent memory. Presented at the NeurIPS (gaze meets ML workshop). https://proceedings.mlr.press/v210/thompson23a.html (P4142)

Greydanus, S., 2023. Scaling down Deep Learning. URL https://greydanus.github.io/2020/12/01/scaling-down/