Construction and disruption of hippocampal cognitive graphs in human and machine

Deng Pan (deng.pan@psy.ox.ac.uk)
University of Oxford, Oxford, UK

Simone D'Ambrogio
University of Oxford, Oxford, UK

Naomi Kingston
University of Oxford, Oxford, UK

Miruna Rascu
University of Oxford, Oxford, UK

Pranav Sankhe
University of Oxford, Oxford, UK

Shuyi Luo
University of Oxford, Oxford, UK

Miriam C. Klein-Flügge
University of Oxford, Oxford, UK

Ali Mahmoodi
University of Oxford, Oxford, UK

Matthew F.S. Rushworth (matthew.rushworth@psy.ox.ac.uk)
University of Oxford, Oxford, UK

Abstract

Humans build internal cognitive graphs that encode the structural relationships between states, goals, and concepts, supporting flexible behaviour. Both the hippocampus (HPC) and orbitofrontal cortex (OFC) are implicated in cognitive map formation, but their distinct roles remain debated. We hypothesised that the HC encodes relational structure among states ("state–state" associations), while the OFC links each state to their goals ("state–goal" associations). To test this, participants performed a structure reversal learning task during fMRI, requiring adaptation to changing state transitions and goals. Computational modelling showed that participants utilised abstract structural knowledge for inference, and multivariate analyses revealed complementary neural representations: the HPC represented the transition structures while the OFC encoded goals. A recurrent neural network (RNN) trained via meta-reinforcement learning (meta-RL) recapitulated these patterns. Disrupting the human HPC using transcranial ultrasound stimulation (TUS) or lesioning HPC-like units in the RNN selectively impaired transition structure learning. Together, these results demonstrate complementary roles: the HPC constructs the foundation of cognitive graphs, while the OFC uses them to support goal-directed behaviour.

Keywords: Hippocampus, Cognitive graphs, Flexible behaviour, fMRI, Meta-RL, Ultrasound stimulation

Introduction

Humans excel at extracting abstract structures from experiences. Like animals constructing cognitive maps of physical space (Tolman, 1948; O'Keefe, 1978), humans extend these representations to organise conceptual knowledge, like complex social relations, forming cognitive graphs (Behrens et al., 2018; Niv, 2019). Such cognitive graphs encode the relational structure between different states and support flexible decisions. Both HPC and OFC are implicated in cognitive map representations, but their specific contributions remain under debate (Rushworth et al., 2011; Wilson et al., 2014; Schuck et al, 2016; Schapiro et al., 2016; Klein-Flügge et al., 2019, 2022). In rodents, HPC is more involved in spatial navigation, while OFC is more related to learning goals or outcomes (Wikenheiser, & Schoenbaum, 2016), hinting at a division of labour that may extend to humans. We hypothesised that the HPC and OFC play complementary roles in building cognitive graphs: HPC primarily encodes relations among states, building the foundation of the cognitive graph, while OFC links states to goals and uses the graph to guide goal-directed behaviour.

Results

Experimental task and behavioural results. We tested these hypotheses using a *Goalkeeper Game* with fMRI, where participants (*N*=36) predicted the shooting direction from four virtual shooters appearing in a probabilistic sequence. In each trial, participants made an initial prediction before seeing the shooter, then could either persist or change their prediction after the shooter appeared. Outcomes provided rewards (+2 correct persistence; 0 correct change) or penalties (−2 any incorrect prediction), incentivising accurate initial predictions. Shooter sequences followed a basic transition structure: a four-state Markov chain with two frequent (paired; 80%) and two infrequent (rare; 20%) transitions. Participants implicitly learned this basic structure with three different character sets during *Pre-Game Training*. Throughout the main game, participants encountered unsignaled reversals: the pair and rare transitions were reversed (transition reversal) twice, and two shooters swapped goal direction preferences (goal reversal) once. Participants successfully adapted, persisting with initial predictions significantly more during paired than rare transitions, indicating that they learned the transition structure and followed an optimal strategy They also accurately learned each shooter's goal preference. These results highlight people integrate state transitions and goals into cognitive graphs to support flexible decision-making.

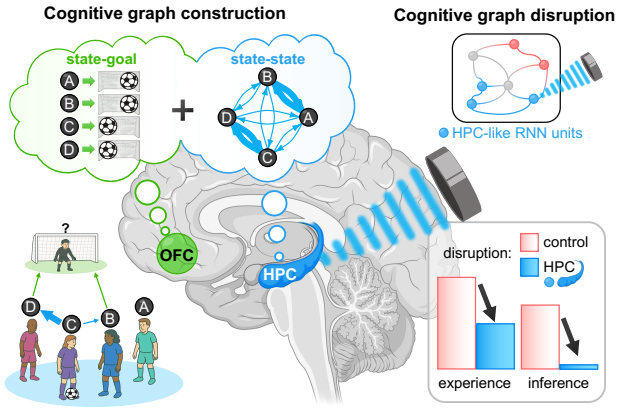
Computational modelling of transition learning. Participants learned the transitions through history, with more recent pair transitions increasing the likelihood of optimal behaviour. Critically, they leveraged both direct experience (same-pair history) and inference (other-pair history) to learn transitions and guide future predictions. We formalised these behaviours with computational models. Three RL models were implemented: a *Frequency Model* (0), which predicts based on recent state (character) frequencies; a *Transition Model* (1), which learns transition probabilities simply through direct experience; and a *Structure Model*, which incorporates structural knowledge to utilise both experience and structure-based inference for predictions. The Structure

model provided the best fit for participants' behaviours and was the only model to replicate participants' use of structure-based inferences, outperforming the other simpler models. This indicates that human learners use structural knowledge to build cognitive graphs and generalise beyond experiences.

Neural representation in HPC vs OFC. fMRI analyses revealed a dissociation between HPC and OFC in representing the cognitive graph. Using trial-level representational similarity analysis (RSA), we found that only HPC consistently represented the transition structure across different phases, with no comparable effects in other task-relevant regions. Paired characters were represented more similarly in the neural space of HPC compared to unpaired characters. Moreover, HPC's representation was sensitive to the global context—it distinguished between different transition structures after reversals, suggesting that it encoded not just local transitions but also a higher-level abstract structure. In contrast, decoding analyses showed that OFC represented the goal direction prediction, and such a representation is disentangled from the transition knowledge. In short, we identified complementary OFC and HPC roles: HPC represents the state-state transition structure to construct the foundation of the cognitive graph, while OFC links states to goals to use the graph to predict goals.

The causal role of HPC in transition learning using TUS neuromodulation. To test the HPC's causal role in cognitive graph formation, we modulated its activity using transcranial ultrasound stimulation (TUS). A new group of participants (*N*=20) completed two counterbalanced sessions, receiving bilateral theta-burst TUS targeting either HPC or a control white matter site after *Pre-Game Training* on the basic transition structure, and then performed the *Goalkeeper Game*. HPC stimulation disrupted participants' transition learning compared to control stimulation. Specifically, it impaired the influence of both experience and structure-based inference, with inference being abolished, suggesting a shift from using structural knowledge (*Model 2*) to a simpler experience-based (*Model 1*) strategy. Notably, participants' goal prediction learning from shot history was unaffected. These findings demonstrate HPC's causal role in learning state-state transitions to construct cognitive graphs and enable flexible behaviour.

Simulating TUS effects with neural network. We next asked whether a similar causal role of structure representation in transition learning could be observed in an artificial network. We trained an RNN via meta-reinforcement learning (meta-RL) with the advantage actor-critic (A2C) algorithm on a state-prediction, following the same transition structure as the *Goalkeeper Game*. After training, the network learned to predict upcoming states using both direct experience and structural inference, closely paralleling human behaviour. Applying RSA to individual hidden units, like the human fMRI analysis, we discovered that about half of the units developed HPC-like representations of the transition structure. Randomly disabling HPC-like units during task performance impaired both use of experience and structure-based inference in proportion to lesion extent, while disabling control units had no effect. Strikingly, lesioning 50% of HPC-like units closely mimicked the effects of hippocampal TUS in humans, highlighting HPC's critical role in cognitive graph construction and offering computational insights into neuromodulation mechanisms.



Conclusion

Using fMRI, RL models, TUS neuromodulation, and RNN simulations, we revealed the complementary roles of HPC and OFC in representing cognitive graphs: HPC encodes state-state transition structure, constructing the foundation of cognitive graphs, while OFC uses the graph for goal prediction. Disrupting HPC representation in either humans or artificial agents impaired structure-based learning, highlighting HPC's pivotal role in cognitive graph construction and supporting structure-based inference.

Key references

Behrens, T. E et al. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490-509.

Klein-Flügge, M. C., Wittmann, M. K., Shpektor, A., Jensen, D. E., & Rushworth, M. F. (2019). Multiple associative structures created by reinforcement and incidental statistical learning mechanisms. *Nature communications*, *10*(1), 4835.

Klein-Flügge, M. C., Bongioanni, A., & Rushworth, M. F. (2022). Medial and orbital frontal cortex in decision-making and flexible behavior. *Neuron*, *110*(17), 2743-2770.

Niv, Y. (2019). Learning task-state representations. *Nature neuroscience*, *22*(10), 1544-1553.

O'Keefe, J. (1978). The hippocampus as a cognitive map.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6), 1054-1069.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3-8.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402-1412.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267-279.

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513-523.