**Title** Neural Mechanisms of Hierarchical Planning during Navigation

**Authors** Balaguer J 1, Hassabis D 2, Spiers H 2, Summerfield C 1

**Affiliations** 1 Dept. Experimental Psychology, University of Oxford, Oxford, UK.

2 Dept. Psychology, University College London, London, UK.

Reinforcement Learning (RL) is a general framework for Machine Learning that also describes mammalian behaviour [1]. In order to allow decisions in complex environments, RL has been extended with the use of o*ptions* [2] that allow agents to learn the value of multicomponential sequences of actions in a given context [3, 4]. Options guide the agent to unique states designated as subgoals, at which a new option is selected. For example, when moving around a building, agents may learn the value of reaching a doorway to enter a new room. However, although prediction error signals for interim goals have been observed in the striatum [5], it remains unknown whether unique representations of subgoal states exist in the human brain.

To address this question, we used fMRI to measure BOLD signal of 20 healthy participants (10 female; age 19-34, mean 25.6 years) performed a navigation task within a previously-learned virtual environment representing a subway network. Each network consisted of multiple stations identified by unambiguous names and connected via "subway lines" that (during training) were represented with a salient contextual cue (colour). We reasoned that stations where the lines intersected (exchange stations) would be treated as subgoals and predicted that they would garner unique behavioural and neural effects, even though during task performance they were not associated with explicit visual cues.

Participants first performed a training session, during which they became familiar with the subway map by repeatedly completing journeys to a destination station. They executed their plan by pressing a key to move in one of the cardinal directions (*north*, *south*, *west*, *east*) at each station, and an additional action (*switch*) to change between lines at an exchange station. Finally, once they reached the destination station, participants received feedback on the duration of their journey (and the optimal distance). During training only, participants could see a map of the network, and were quizzed about it to ensure that they were learning.

On a subsequent day, participants undertook a second session whilst undergoing whole-brain fMRI. In the scanner they also completed multiple journeys but the colour of the lines were removed and no additional action was required to switch between lines. . Participants earned bonus income by reaching their goal on time, and thus tried to minimise the absolute distance of their paths. Journeys were randomly cancelled with fixed probability on each time step, to put pressure on participants to decorrelate key experimental variables for imaging anlaysis.

Reaction times increased as the subject approached a line change, and then speeded up directly after the change. Critically, this did not happen around exchange stations where the subject did not change lines.

On the neural level, we observed higher activation in the bilateral dorsolateral prefrontal cortex when an exchange station was reached, which was independent of a line change, suggestive of processing the demand associated with changing context (or not). An increase of activation in the motor and premotor cortices was observed for a switch of direction (i.e. response) from the previous trial, independently of being in an exchange station or not. Specifically for line-changes, i.e. when the response was changed *and* it was an exchange station, we found higher activation in a network of right caudate nucleus, thalamus, and the supplementary motor area. This suggests that the striatum signals the need to switch to a new context at a subgoal state. Meanwhile, acitivty in medial Prefrontal Cortex (mPFC) increased as subjects approached the final destination, as if it were encoding proximity to a goal state.

This results provide evidence that a unique network is activated when participants reach an interim goal during navigation. Our findings suggest that humans represent the value of actions in a hierarchical fashion, and that the interconnected structures signal when a subgoal has been reached.

1. Sutton, Barto (1998), *Reinforcement Learning: An Introduction*, The MIT Press
2. Sutton, Precup, Singh (1999), *Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning*, Artificial Intelligence, vol. 112, Issues 1-2, pp. 181-211
3. Schapiro et al. (2013) *Neural representations of events arise from temporal community structure*, Nature Neuroscience, 16, pp. 486–492
4. Botvinick, Niv, Barto (2009), *Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective,* Cognition, vol. 113, Issue 3, pp. 262-280
5. Ribas-Fernandes et al. (2011) *A Neural Signature of Hierarchical Reinforcement Learning*, Neuron, vol. 71, Issue 2, pp. 370-379