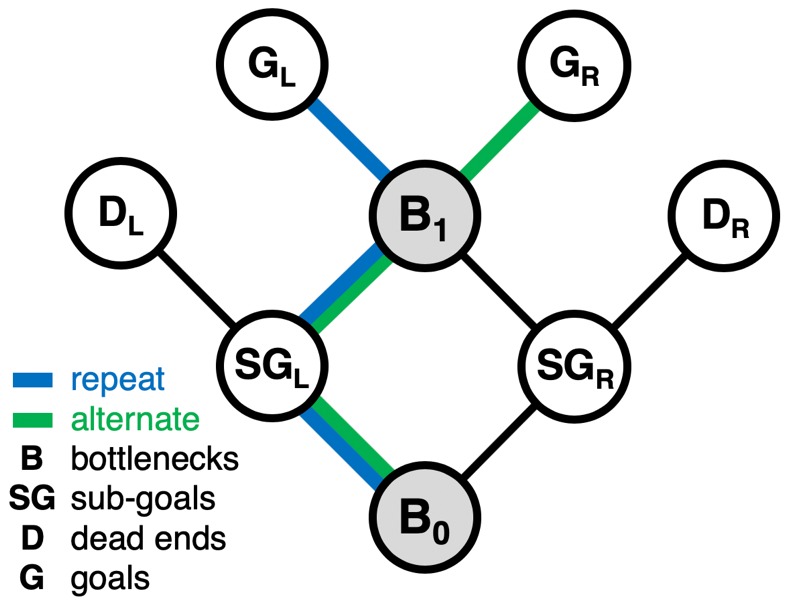
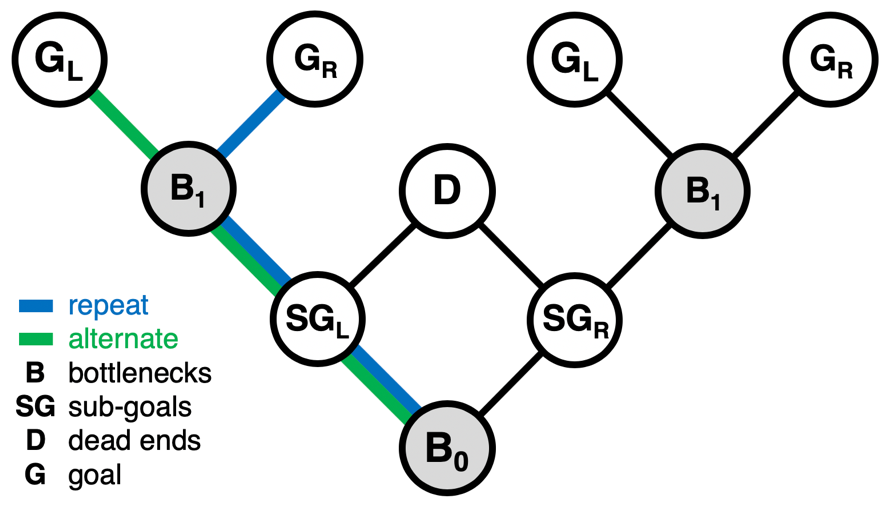
The pervasively hierarchical structure of human (and non-human) behaviour requires learning and decision-making algorithms able to engage the issue of hierarchy. Hierarchical reinforcement learning (RL), which leverages temporal abstraction (i.e., the representation of temporally-extended macro-actions), is one such algorithm. It addresses the scaling problem present in flat (i.e., non-hierarchical) RL methods, is capable of furnishing flexible and efficient systems, and could plausibly be employed by the brain (Botvinick, Niv & Barto, 2009). One popular and influential implementation of hierarchical RL is the options framework (Sutton, Precup & Singh, 1999), which supplements the set of primitive actions present in flat RL with temporally abstract options. Whilst there is a growing body of evidence supporting the relevance of the options framework for the human brain (Ribas-Fernandes, Solway, Diuk, McGuire, Barto, Niv & Botvinick, 2011;Diuk, Tsai, Wallis, Botvinick & Niv, 2013), there is as of yet only indirect behavioural evidence supporting options as relevant to the hierarchical organisation of behaviour observable in humans.

I propose to build a set of reinforcement learning agents and to have them complete a navigation task (see figure 1 and its heading for a detailed description of the task). Subjects must learn to navigate through the environment in line with one of two regimes towards one of two goal locations. Once performance has stabilised, the to-be-followed regime changes. Humans are able to learn to solve the task with near-perfect accuracy and can learn to behave in line with a new regime following a switch after only one trial under the new regime; people exhibit one-shot learning of the novel experienced path and zero-shot learning of the novel unexperienced path. I propose to investigate how different types of RL models may be able to reproduce this ability to learn and generalise from minimal experience and to compare the behaviour of these models in different contexts in order to delineate their behaviour and guide further experiments.

**Figure 1** – The map of all possible state transitions in the task. The agent starts at B0 and must navigate to either GL or GR (this will be instructed on each trial by some input). In order to receive reward at the appropriate G location, the agent must get there via a given sub-goal location. The correct SG location is defined by a combination of the current goal location and the imposed regime of the current trial (repeat/alternate). For example, if the goal location is GL and the regime of the current trial is repeat, the appropriate sub-goal location is SGL. The agent will receive reward at GL only if it traverses SGL on its way there (i.e., if it takes the blue path shown in the figure). Reward will be placed with even probability at GL and GR, meaning that . Solving the task thus requires some form of action hierarchy or state representations that include history. Once the agent has learned to solve the task, the imposed regime will switch, and the agent will need to recover by learning to behave in line with the new regime.

**For more detail, keep reading, though most of this next section needs some more guidance…**

I propose to start by building a model-free hierarchical reinforcement learning agent under the options framework with pre-defined options that allow the agent to solve the task. Having done so (and having developed my own experience with working with these models), I plan to develop the agent such that it is able to learn options through experience with its environment. This requires tackling the option discovery problem, for which several different methods have been proposed in the computational RL literature. I also aim to enhance the option representations held by the agent such that they are able to be general (i.e., such that a single option is able to define repetition at B0 and B1 independently of the specific primitive action being repeated). Given a sufficiently fast learning rate, I would expect this final model to exhibit both the one-shot and zero-shot learning observed in humans. For comparison, I plan to build flat model-free RL agents that hold state representations that either do or do not include representations of history. Without history representation, the agent should be unable to perform above chance, whilst with history, the agent should learn to optimally perform the task. More interesting and productive comparisons can be made between the competent flat and hierarchical agents (i.e., which is faster to learn, which is faster to generalise, how does each behave if placed in a structurally similar but different environment such as that shown in figure 2). Delineating the behaviour of these two agents (and any others that are able to perform the task and qualitatively capture human behaviour) could guide experiments aiming to provide direct behavioural evidence pertaining to the relevance of the options framework for the hierarchical organisation of human behaviour.

**Figure 2** – An alternative state map which manipulates the accuracy of any low-level options that guide behaviour between B0 and B1. In the map shown in Figure 1, two useful low-level sequences of behaviour are the left-right and right-left transitions required to move from B0 to B1 (via SGL and SGR respectively). Should a hierarchical agent learn to represent and use these options, it would make errors at SG locations if placed in this environment but would still respond correctly in the following bottleneck B1. Patterns of errors, transfer, and generalisation during recovery when introduced to this environment may delineate between the outlined agents.

References

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behaviour and its

neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262-280.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces

two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, *33*(13), 5797-5805.

Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., &

Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370-379.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework

for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181-211.