Assembling hierarchies of action using sequencing and abstraction: studies and models of zero-shot learning

Gwydion Williams (gwydion.williams.15@ucl.ac.uk)
University College London
Stefano Palminteri
École Normale Supérieure https://orcid.org/0000-0001-5768-6646
Patrick Haggard
University College London https://orcid.org/0000-0001-7798-793X

Article

Keywords:

Posted Date: February 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1312094/v1

Assembling hierarchies of action using sequencing and abstraction: studies and models of zero-shot learning

Gwydion Williams[1], Stefano Palminteri[2,3,4], Patrick Haggard[1]

[1] Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AZ, UK

[2] Laboratoire de Neurosciences Cognitives et Computationnelles, INSERM, Paris, France

[3] Département d'Etudes Cognitives, ENS, PSL Research University, Paris, France

[4] Institute of Cognitive Neuroscience, HSE, Moscow, Federation of Russia

Corresponding author: Gwydion Williams; gwydion.williams.15@ucl.ac.uk

Abstract

Although the hierarchical structure of human action is widely acknowledged, we do not fully understand how hierarchies of action are assembled. The standard view is that low-level actions are *sequenced* to establish higher-level routines of behaviour. Here we develop an alternative approach to building hierarchies, based on two insights. First, we consider relations between sequence elements. Second, we identify *abstract* features common to several such relations, and show how these abstract features allow for flexible action sequence learning. We combine *sequencing* and *abstraction* within a single model of hierarchical structure and test this model in two distinct versions of a novel experimental paradigm. We demonstrate that humans can learn entirely novel sequences of actions without practice, by generalising learned sequence structures from one context to another. Computational modelling showed that this 'zero-shot learning' of novel behaviours was successfully captured by a hierarchical organisation of the kind we propose.

Introduction

As Lashley (1951) seminally observed, all actions we produce are component parts of some sequence, and sequences of action are best understood through a hierarchical lens. Learning to sequence actions, and then to flexibly arrange these sequences into higher-level routines, is essential for many everyday tasks (Rosenbaum, Kenny, & Derr, 1983; Yokoi & Diedrichsen, 2019). For example, consider the hierarchy of behaviours required to brew coffee. To begin, we must fetch a mug, some ground coffee, and a kettle of boiled water. To boil a kettle, we must lift and move it to a tap, fill it with water, return it to its power source, and turn it on. Each of these elements can in turn be further decomposed: to lift a kettle, we must locate its handle in space, reach for the handle, and grasp it. This cascade of ever lower-level representations of action establishes an entire hierarchy of behaviours that, taken as a whole, will satisfy the original goal of making coffee (see Figure 1). A hierarchical organisation is most often justified by its computational efficiency; limited cognitive capacity can be dedicated to high-level features of action, while low-level details are delegated to modular circuits at lower levels. In this study we provide evidence and argue for a second benefit of hierarchy: it does not only minimise cost, but it also maximises benefit by speeding rule discovery.

Classically, sequential action has been described as being a process of building up chunks of behaviour by sequencing elementary or primitive actions (Lashley, 1951). Under the classical view, each chunk of action would activate its motor components in order and chunks themselves could be sequenced to form progressively higher-level routines of action. This hierarchical sequencing of lower-level actions to produce higher-level representations of order facilitates faster and more accurate execution of primitive actions (Rosenbaum et al., 1983), provides a more computationally efficient scheme to store and recall sequences of behaviour (Ramkumar et al., 2016), and allows for entirely new sequences to be learned by combining existing chunks in novel orders (Sakai, Kitaguchi, & Hikosaka, 2003). Hierarchical sequencing has been observed in the study of sequential motor control in humans (Cooper & Shallice, 2000; Fuster, 2008; Humphreys & Forde, 1998; Miller, Galanter, & Pribram, 2017; Yokoi & Diedrichsen, 2019), and it is the organisational principle used to arrange actions in hierarchical reinforcement learning (see *temporal abstraction*: Botvinick, Niv, & Barto, 2009; Botvinick, Weinstein, Solway, & Barto, 2015; Solway et al., 2014; Sutton, Precup, & Singh, 1999). According to this sequencing of low-level parts, high-level representations of sequence therefore lack fine temporal detail of the low-level movements involved, and they need only store information of the order in which constituent actions must be initiated.

There is evidence for a second and more abstract mode of representation at high levels; single neuron (Shima, Isoda, Mushiake, & Tanji, 2007) and population (neuroimaging) data (Kornysheva et al., 2019) converge on the notion that, at high levels, the relations between sequence elements are represented independently of the elements themselves (see *abstraction* in Figure 1). These findings suggest that

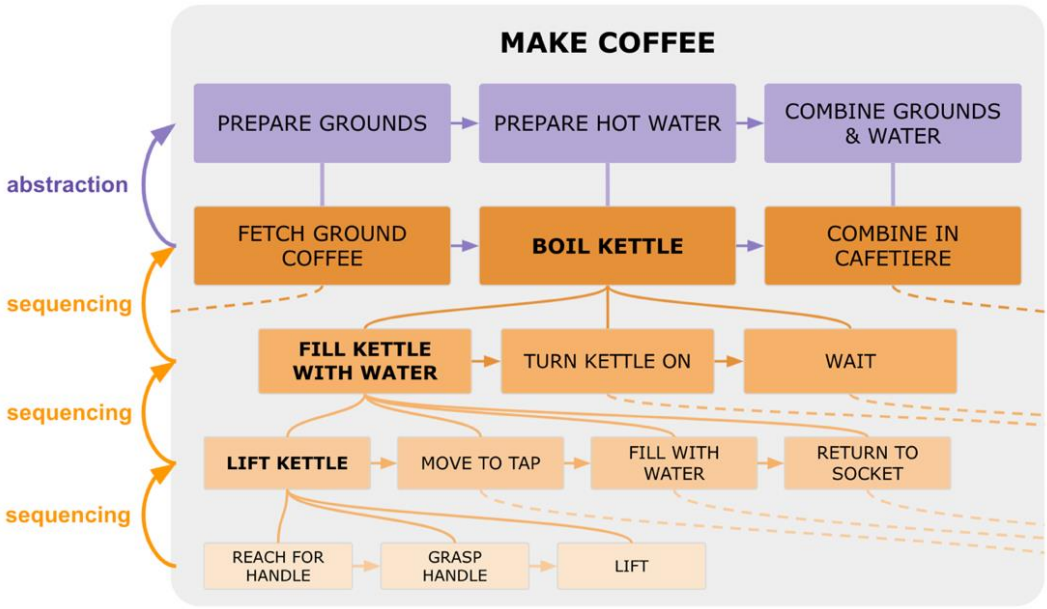

Figure 1 – Hierarchy of actions required to make coffee. Higher-level representations of action can come from two distinct operations: (1) sequencing low-level actions (e.g., reach for and grasp the handle of a kettle) can provide higher-level representations (e.g., lift kettle); and (2) abstracting over the individual actions in a sequence can provide abstract and relational representations of the relations between sequence elements independent of their content. This second method of abstraction can allow for the same relational representation (in purple) to produce distinctly different low-level sequences that adhere to the same relational structure (e.g., fetch ground coffee could be replaced with grind coffee beans to satisfy prepare grounds).

the human brain not only abstracts away the fine details of a motor command, but in some cases the actions themselves are lost in favour of a representation of the *relational structure* of the sequence (e.g., whether a given action should be repeated or not) independently of the specific action elements involved. Humans may use both *sequencing* and *abstraction* to form high-level representations of action, but it is unclear whether these two distinct operations are integrated under a single hierarchical framework to control behaviour, and the benefits doing so have not been explored.

We propose that *sequencing* and *abstraction* (see Figure 1) as two methods of building up higher-level routines of behaviour from lower-level actions are combined by the human brain under a single hierarchical organisation of behaviour. We further propose that this combination establishes an efficient, generalisable, and adaptive structure of human action. How might we detect this high-level organisation from low-level behavioural data? Movement patterns are typically silent about the generative processes that produce them. Further, observable movements represent the direct output of low-level modules, and recovering underlying higher-level structure is difficult because it is filtered by lower-level modules. Here we propose a new approach to extracting abstract hierarchical representations from behavioural data based on immediate generalisation of learned sequence structure to produce entirely novel action sequences that meet completely new challenges. We refer to this process as *zero-shot learning of novel behaviours*. We reasoned that, if people indeed form relational representations during learning complex action sequences, this should allow immediate generalisation to new action sequences that share the same relational properties but involve distinct low-level actions. For example, consider the abstract representation of the steps required to brew coffee in Figure 1. If one holds this abstract and relational representation of the steps required to brew

coffee, then when faced with a new coffee maker (say, a filter coffee machine), one may be able to learn quickly how to brew coffee with the new apparatus by using this abstract high-level representation of the steps required to produce an entirely novel sequence of low-level actions. We propose this abstraction and generalisation of structure to produce novel behaviours as a novel behavioural marker of latent hierarchical structure.

In the present study, we demonstrate that humans do indeed exhibit zero-shot learning of novel behaviours. We report two experiments on goal-directed action which use very different visual presentations and framing, but an identical underlying structure. To earn reward in the tasks, participants needed to navigate to a sub-goal before moving on to an end goal. For example, in one of the two tasks, subjects would need to find a key (sub-goal) in one room before opening a chest (goal) in another room. There were two possible sub-goal locations and two possible goal locations, but only one of each was active on a given trial (i.e., the key would be placed in one of two rooms on each trial, as would the chest). The locations of the sub-goal and goal were associated, such that one could predict the location of the goal from the location of the sub-goal. Importantly, participants were told only where the sub-goal was located on each trial, and the central challenge was to learn to predict the location of the goal from that information, so as to achieve reward in as few steps as possible. Given that there were multiple different sub-goal locations, any single association between sub-goal and goal could require distinct sequences of action on different trials. Further, these sub-goal-to-goal associations could change without warning. We observed that participants learned new associations from only a single trial following a switch. Crucially, they also immediately generalised knowledge of the new association to produce entirely novel sequences of low-level actions on subsequent trials. To formally verify that this zero-shot learning of novel behaviours is indeed evidence of a hierarchical system that included both sequencing and abstraction, we used computational modelling to explore the necessary cognitive components of this learning process. We found that we could only replicate zero-shot learning with a system that (1) organised behaviour hierarchically by sequencing lower-level parts to provide higher-level representations of ordered elements, (2) made use of relational high-level representations of action by *abstracting* over lower-level sequences, (3) abstracted learning about these relational representations over multiple states, and (4) directed exploration at appropriate hierarchical levels. In sum, we present novel findings and related computational evidence showing that hierarchical reinforcement learning is useful not only for an efficient storage of wide repertoires of behaviour, but also for impressively fast adaptation to changes in the environment when it is combined with highly abstract representations of action.

Results

Behavioural protocol and task structure.

Our behavioural paradigm sought evidence for specific hierarchical representations that specify the relations between actions within a sequence. Participants were to navigate around the state map seen in Figure 2A in search of a sub-goal location (SG on the map). Visiting the subgoal would then allow them to receive reward at a separate goal location (G on the map). We used this state map to build two tasks which appeared to be very different (see supplementary materials) but were in fact structurally identical. In a spatial version of the task, participants navigated a set of rooms in search of a key (SG) that would open a chest (G). In a procedural version of the task, participants solved a puzzle by moving a rod to a specific cube-face (SG), which would then unlock reward at another cube-face (G). At subsequent debriefing, none of the participants reported recognising any similarities between the two tasks despite their identical structure.

The state map underlying both tasks was designed to require a specific hierarchy of actions to navigate efficiently around it (see Figure 3 for the full

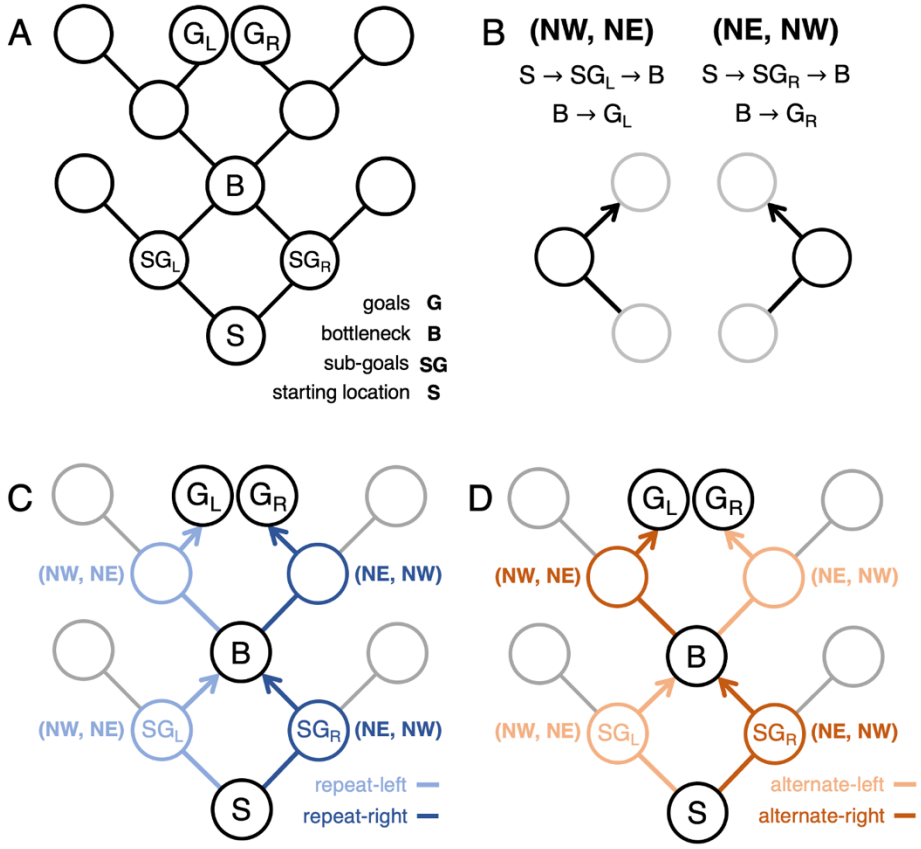

Figure 2 – (A) map of state space followed by both spatial and procedural tasks; (B) illustration of useful chunk of actions for navigating to/from the bottleneck state; (C) illustration of the two distinct sequences of action required if the association between SG and G is *repeat*; (D) illustration of the two distinct sequences of action required if the association between SG and G is *alternate*.

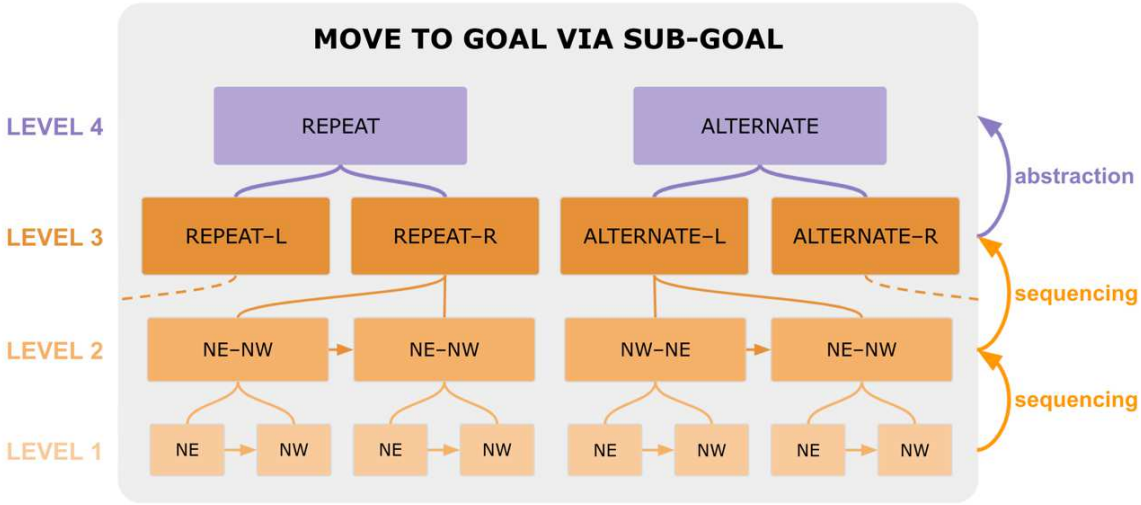

Figure 3 – Schematic of the hierarchy of actions targeted by the task design. Level 1 comprises the four primitive actions available in the task. Level 2 contains length-2 sequences that are useful for navigating to/from the central bottleneck (see Figure 1). Level 3 contains the full sequences of action required for an optimal solution of the four possible trial types (i.e., for all combinations of sub-goal–goal associations and sub-goal locations); and finally at level 4 we find abstractions over the two sub-goal–goal associations towards a relational representation of the actions involved.

hierarchy). The bottleneck in the centre of the map (see Figure 2A) needed to be traversed on all trials, and it needed to be traversed to move from the bottom half of the space to the top half, making it a useful target for behaviour. From the start position (S), either a sequence of (NW, NE) or a sequence of (NE, NW) would move participants from the starting location to the bottleneck (see Figure 2B). Given the symmetry between the bottom and top halves of the map, these same sequences were sufficient to then move from the bottleneck to each of the two possible goal locations. The four primitive actions (NE, NW, SE, SW) therefore occupy the lowest level (*level 1* in Figure 3) of our target behavioural hierarchy, and the chunks of 2 sequential actions that are used for travelling to and from the bottleneck are one hierarchical level above the primitive actions (*level 2* in Figure 3; see Figure 2B for a demonstration).

On a given trial, only one of the two sub-goals (SGL or SGR) and one of the two goals (GL or GR) was active. For example, in the spatial task the participant would discover a key in only one of the two sub-goal rooms and a chest in only one of the two goal rooms. Participants were told at the start of a trial which of the two sub-goal states they should visit, and this therefore guided which of the two *level 2* sequences they should execute (see Figure 2B). Importantly, participants were not told which of the two goal locations was active, but the location of the goal could be predicted from the location of the sub-goal. Participants were told that they could predict where the goal would be from where the sub-goal was, but they were not told how to make this prediction. There were two possible associations between sub-goal and goal: (1) the goal could be on the same side as the sub-goal, or (2) the goal and sub-goals could be on different sides. We refer to the first of these two associations as *repeat*, and second as *alternate*. If a participant selected the correct *level 2* sequence such that they travelled to the bottleneck via the active sub-goal, then upon reaching the bottleneck they would need to decide between repeating the *level 2* sequence that got them there or alternating to execute the other of the two *level 2* sequences. The correct decision here would depend on the current association between sub-goal and goal: if the association was *repeat*, then the correct decision

is to repeat whatever *level 2* sequence used to reach the bottleneck, and if the association is *alternate*, then one should alternate. This repetition of or alternation between *level 2* sequences establishes four higher-level representations of the sequences of actions required to solve the task (*level 3* in Figure 3): there are two repetition sequences (one each for travelling via the left and right sub-goal, see Figure 2C), and two alternation sequences (again, one each for travelling via the left and right sub-goal, see Figure 2D). Finally, there is potential for an abstraction over the *level 2* sequences being repeated in *level 3* such that our participants would represent *repetition* and *alternation* independently of the *level 2* action sequences being repeated or alternated (*level 4* in Figure 3). Crucially, participants were never explicitly told whether they should repeat or alternate, but they could derive this information by correctly learning and representing the relation between the sub-goal and the goal, i.e., by representing the hierarchical and relational structure of the task.

The tasks were organised into three blocks of at least 30 trials each. In the first block, the sub-goal-to-goal association was fixed. In the two blocks that followed, the association between sub-goal and goal would switch on one of the first 10 trials, and participants would then complete 30 trials under the new association (see Figure 4A). We refer to the trials where these switches in association occur as *switch trials*. Participants were informed in the instructions that the associations between sub-goal and goal could occasionally change. A switch trial could occur on a trial where the sub-goal was present in either of the right or left locations, and so participants first experienced the new association along only one of two possible paths through the environment. For example, one participant might have experienced a switch from *repeat* to *alternate* on a trial where the sub-goal was on the right, and they could then learn how to act under *alternate* when the sub-goal is on the right. When the sub-goal is next on the left, although the sequence of actions required will adhere to the same *alternate* structure learned via the right, it requires a completely novel sequence of low-level actions (compare the two *alternate* paths in Figure 2D). We refer to the first trial along this inexperienced path following a switch in sub-goal-to-goal associations the *novel post-switch trial*, and we refer to the novel sequences of actions required on these trials as the *novel paths*. Given that the sub-goal is randomly allocated to the right or left trial by trial, the novel post-switch trial might not necessarily follow immediately after the switch: in our dataset the maximum number of trials between a switch trial and its associated novel post-switch trial was four).

Immediate Acquisition of Novel Sequences

On both spatial and procedural tasks, all participants learned within the first nine trials how to travel to the correct goal via the active sub-goal in an optimal four moves (for the spatial task, median number of trials taken to make the optimal four moves to goal was 3.5, inter-quartile range = 6.25; for the procedural task, median = 4.5, inter-quartile range = 2.75). Learning was slightly slower on the procedural task (see the shallower rate of learning in Figure 4B), though behaviour did nevertheless converge on the optimum of four moves to goal. The slower rate of learning on the procedural task (learning rates found by fitting exponential models of learning to each participant's data for each domain were significantly slower for procedural than

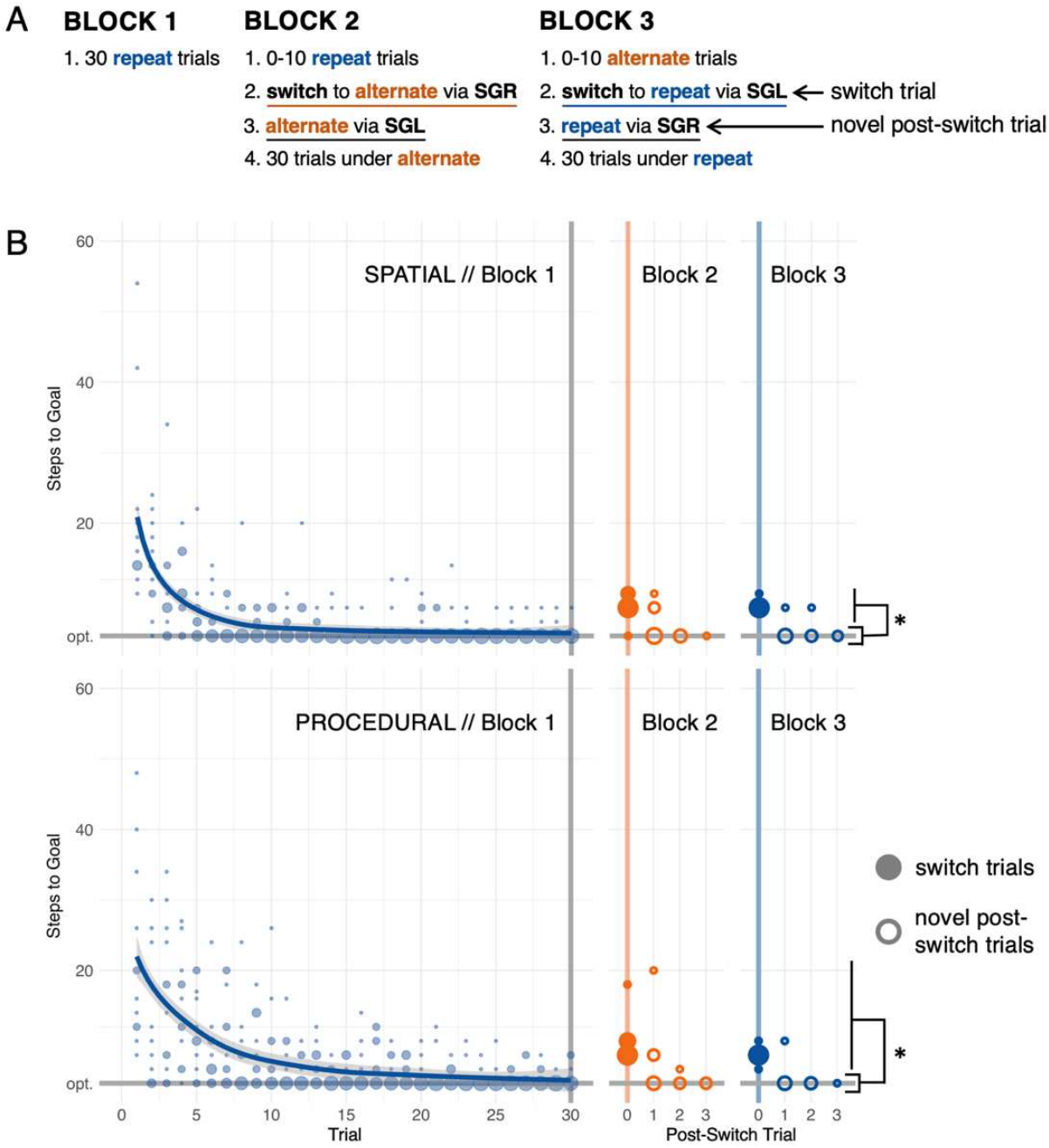

Figure 4 – (A) An example of the procedure followed by each of the tasks (note that the order of SG-G associations was counterbalanced over participants); (B) Observed behaviour of 12 subjects on each of the spatial and procedural tasks. The first column plots behaviour of all 12 subjects in the first block of each task to demonstrate an initial phase of learning and an eventual convergence onto the optimal solution to both tasks. The following two columns present recovery after a switch in SG-G association. The vertical orange/blue bars are the switch trials (these correspond to the underlined switch trials in A), and the hollowed-out points that follow plot behaviour on the novel post-switch trial for all twelve participants (these correspond to the underlined novel post-switch trials in A). Across the board, for any number of trials in between the switch and novel post-switch trials, participants were more likely than not to exhibit optimal behaviour, and this was true in both spatial and procedural tasks.

for spatial, $t(11) = 2.61, p = .024, d = 0.75$) may be due to the unfamiliar setting. Once participants found the optimal solution, they generally continued to perform optimally (see the stable optimal behaviour in block 1 of Figure 4B), with only minor and infrequent deviations, presumably reflecting lapses in attention.

Our central interest here was in how quickly our participants could recover from a switch in the associations between sub-goal and goal. Specifically, we wanted

to ask whether participants would behave optimally on novel post-switch trials despite having no experience of travelling along the corresponding novel path. That is, we were searching for zero-shot learning. This would require a high-level relational representation of alternation and repetition (as in *level 4* of Figure 3) that participants could use to adaptively generate completely novel sequences of behaviour that followed these relational structures. For example, participants could learn that alternating via the left sub-goal following the switch would mean that they should also alternate via the right sub-goal, and upon first visiting the right sub-goal they would know immediately how to solve the task. We found that most of our participants selected the optimal path on novel post-switch trials; of a total of 48 novel post-switch trials, behaviour on 37 of these trials was optimal ($\chi^2(1) = 14.08, p < .001$). Further, the proportions of post-switch trials that were optimal for each subject deviated significantly from a conservative chance level of 0.5 ($t(11) = 3.22, p = .008, d = 0.93$). The number of intervening trials in between the switch and novel post-switch trials (see Figure 4) had no significant effect (the number of intervening trials did not predict a significant portion of variance in steps to goal on novel post-switch trials, $F(1, 46) = 0.94, p = .336$). Finally, there was no evidence to support an association between optimality on novel post-switch trials and task domain (results from chi-squared test for association: $\chi^2(1) = 0.12, p = .731$), indicating that the ability to learn immediately how to act was general and not tied to any individual task. That is, our participants spontaneously generalised learned sub-goal–goal associations to produce entirely novel and optimal sequences of behaviour, an observation which we refer to as *zero-shot learning*. Crucially, the level 1 actions on switch and novel post-switch trials are entirely different, which requires an abstraction over the sequences produced on switch trials to later produce a novel sequence of behaviour that follows the same relational structure.

Note that 0.5 is a very conservative chance level for the likelihood of mistakenly performing zero-shot learning of the novel path following a switch. In reality, if our participants understood nothing of the high-level relations between sub-goal and goal, then there would be no reason to think that any change in association between the left sub-goal and its corresponding goal location should result in a change in association between the right sub-goal and its corresponding goal location. As a result, if a switch to *alternate* trial fell on a trial where the sub-goal was on the left, when next encountering a trial where the sub-goal was on the right, the rational choice would be to follow the association that was active before the switch (*repeat*) and not the new association learned via the left sub-goal (*alternate*), making the chance level for *alternating* via the right sub-goal 0. In fact, we found a mean proportion of 0.77 (SD = 0.29) of novel post-switch trials being optimal over our subjects, providing strong evidence for the ability to perform zero-shot learning of novel sequences of action. We planned to derive estimates of the true chance level for zero-shot learning from our simulations, which we discuss in greater detail below.

Computational Models

To illustrate how zero-shot learning of novel behaviours arises from hierarchically organised behaviour that makes use of abstract relational representations of action and to search for any other necessary cognitive components of the process, we built a systematically organised set of four different RL models that aimed to capture our participants' behavioural data (see Table 1 for a summary of differences between the four models). The first and simplest model (Model 1 or *flat-history*) is the only non-

hierarchical model included, meaning it only has access to the four primitive actions (see Figure 3). It makes use of memory to solve the task (which is required given that the task is non-Markovian), where the remaining four models use hierarchically organised action to solve the task. We used standard Q-learning over temporal difference prediction errors (Sutton, 1988; Watkins & Dayan, 1992), and modelled participants' choices using a softmax function. This first model provided a non-hierarchical baseline against which we could compare the performance of our more complex hierarchical RL models.

The three remaining models were all hierarchical. All three follow the options framework (Sutton et al., 1999), which supplements the primitive actions available in standard, flat RL with temporally-abstract *options*, corresponding to superordinate chunks of behaviour. Our three HRL models hold the options required to furnish particular subsets of the behavioural hierarchy outlined in Figure 3. Models 2 (*simple-hierarchical*) and 3 (*structured-hierarchical*) hold the first three levels of the hierarchy, but only model 4 (*abstract hierarchical*) holds the abstract and relational representations of repetition and alternation. Model 4 also abstracts learning over trials where the sub-goal is on the right and trials where the sub-goal is on the left.

We hypothesised that a preference to explore at high rather than low levels is central to the ability to quickly learn and use high-level relational rules, and to test this we implemented a specific modification of the softmax function in models 3 and 4. Whereas standard softmax would include all actions/options no matter their hierarchical level, our structured-softmax chooses between only the highest-level options available given the current state of the agent. Further, where standard softmax uses a temperature parameter that is insensitive to any features of hierarchically organised action, our structured-softmax modifies the value of its temperature parameter as a function of the hierarchical level of the options under consideration. In practice, models 3 and 4 therefore choose only between the highest-level actions available in a given state and choice in general is biased towards exploration at higher hierarchical levels, and exploitation at lower levels.

Table 1 – Key differences between our four models.

	1. Flat w/History	**2. Simple Hierarchical**	**3. Structured Hierarchical**	**4. Abstract Hierarchical**
Levels available (see Figure 3)	1 (primitive actions only)	1-3 (hierarchy with no abstraction)	1-3 (hierarchy with no abstraction)	1-4 (hierarchy with abstraction)
History representation	Learns $Q(a, s_t, \mathrm{SG})$	Implicit in option execution	Implicit in option execution	Implicit in option execution
Hierarchical operations performed	Sequencing	Sequencing	Sequencing	Sequencing; Abstraction
Policy	Softmax	Softmax	Structured-Softmax	Structured-Softmax
Selects between…	All available actions	All available actions	Highest-level available actions	Highest-level available actions
Performs state abstraction?	No	No	No	Yes (between SG-R & SG-L trials)

In the first step of our computational analyses, we used simulations to see to what extent our models could generate zero-shot learning of novel paths, as observed in the behaviour of our participants. In the second step, we fit estimable versions of these models to behaviour to move beyond the few trials where learning of novel paths could take place and to investigate the global process of learning to solve the entire task.

The Necessary Components of Zero-Shot Learning

To estimate how frequently each of our four models could reproduce zero-shot learning by behaving optimally on novel post-switch trials, we simulated the behaviour from each model for a range of parameter values. We manipulated learning rate (alpha) and temperature (beta) to establish a grid of parameter values (each of these two parameters could occupy any of the following values: 0.2, 0.4, 0.6, 0.8, 1.0), and for each combination of learning rate and temperature within this grid we simulated behaviour on the task 20 times. From these simulations, we computed the proportion of novel post-switch trials where behaviour was optimal. We use these simulations not only to investigate the success of our models, but also to estimate the true chance level of producing zero-shot learning under our various models and the hypotheses they test.

We found that only model 4 (abstract-hierarchical) exhibited proportions of zero-shot learning close to those observed empirically. As expected, our non-hierarchical baseline produced almost no zero-shot learning, and this model provides a good estimate of the true chance level of behaving optimally on novel post-switch trials if we make no assumptions about the structure of behaviour. Models 2 (simple-hierarchical) and 3 (structured hierarchical) lead to modest incremental improvements as the organisation of behaviour becomes more sophisticated. However, these two models only produce zero-shot learning by chance; they must explore the options available to them on novel post-switch trials, and should they happen to explore by selecting the newly-optimal high-level routine of action, we would then see zero-shot learning. Model 4 offers a qualitative change in this process, as it is able to learn from one context how to behave in another. That is, model 4 can learn the abstract relations between sub-goal and goal from experience with only one of the two sub-goal locations, and it can apply these learnings to guide behaviour when it next encounters the other sub-goal. Unsurprisingly, therefore, the success of model 4 in capturing zero-shot learning grows monotonically with learning rate (see Figure 5B). A higher learning rate allows the model to learn new abstract relations between sub-goal and goal from only a single trial. In summary, models 1, 2, and 3 fail to capture zero-shot learning of novel behaviours, but model 4 succeeds. Neither hierarchical organisation nor a preference for high-level exploration alone were sufficient to capture zero-shot learning, but when combined with abstract relational representations of action and an ability to abstract learning over distinct states, all four components allowed model 4 to exhibit a near-human ability to generalise learned structure to produce entirely novel sequences of action.

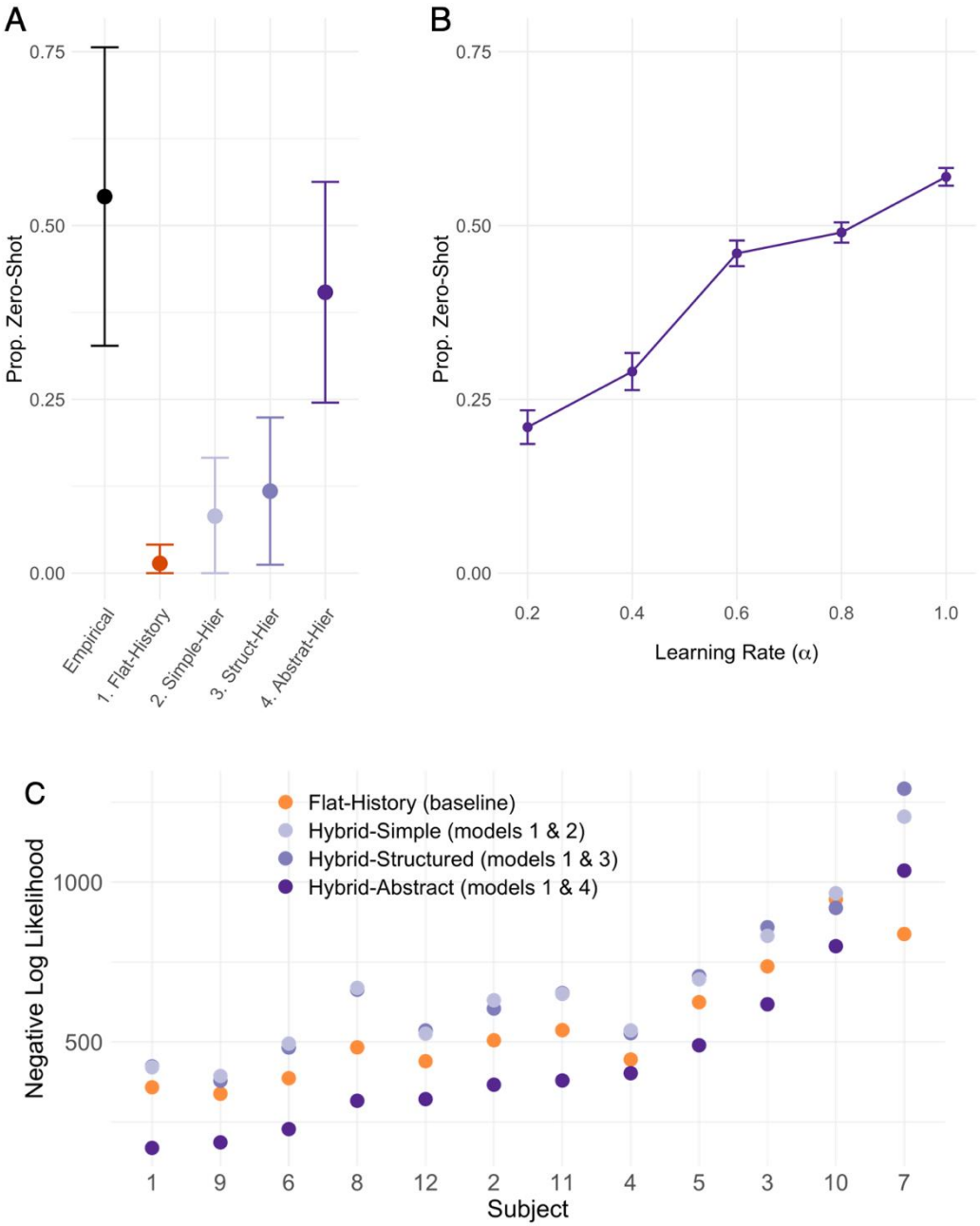

Figure 5 – (A) Mean (± SD) proportion of replications that exhibited zero-shot learning for a range of learning rates and temperatures for all four models, with the empirical means plotted for comparison. We see incremental improvements as we increase the complexity of our hierarchical models, but only model 4 is capable of reaching near-human performance. (B) Plot of how the ability of model 4 (our most successful model from (A) to capture zero-shot learning varies with learning rate – we find a monotonic increase in success with learning rate. (C) Fits of our flat-history (baseline) and three hybrid models to all participants. For 11 of 12 participants, the hybrid-abstract model clearly fits best, with the one remaining participants being fit best by our baseline flat model.

Model Fits to Behaviour

Our models were designed to capture one key behaviour of interest, namely zero-shot learning at the novel post-switch trial. However, zero-shot learning corresponds to a single sequence of actions within a much larger sequence of navigational or problem-solving actions (i.e., the entire task). We therefore additionally fitted these models to behaviour in the task, to investigate their generality, in addition to their local fit. In fact, our hierarchical models learned to use built-in options that were

designed to meet the demands of the task, while in reality the agent first needs to learn from experience with the task what these useful options might be. Practically, this means that our hierarchical models are unable to capture the initial period of learning how to solve the task. We reasoned that this reflects the intuition that an agent in a novel environment must first explore the outcomes of their low-level actions and learn the structure of their environment, and only thereafter can they build a hierarchical structure able to exploit the relational and structural features of the task. We therefore decided to hybridise our models by using model 1 (*flat-history*) again as a non-hierarchical baseline and to additionally combine this non-hierarchical element with each of the hierarchical models in turn, producing three distinct hybrid flat + hierarchical models. To specify the hybridisation process itself, we included an arbitration process that apportioned control of behaviour between flat and hierarchical systems, via an additional parameter, omega. When omega > 0.5, the flat system predominates, while for omega < 0.5 the hierarchical system predominates. The value of omega decays exponentially over time reflecting a shift, with experience, from a flat system of behavioural control to a hierarchical organisation of action. The agent must begin the task with a flat organisation of behaviour (as it does not yet know the structure of the task) but with time discovers a useful hierarchy of actions. The hybridising approach was intended to capture this transition from flat to hierarchical behaviour while still allowing us to compare the performance of each of our hierarchical models against experimental data.

We fit our hybrid models to behaviour using standard maximum likelihood estimation. All four models were fit with only two free parameters – learning rate, and temperature. The hybrid models set omega (governing arbitration between flat and hierarchical systems) and its decay parameter to be fixed at values of 0.9 and 0.95 respectively. Fixing these parameters was necessary, as in order to fit our hybrid model to behaviour, we had to permit occasional errors in behaviour to be attributed to the flat system included in the model whatever the value of omega. In practice, this means that we would occasionally allow the flat system to take control despite the value for omega being below the threshold that would allow this to take place as per the model specification. This slight deviation from the specification was necessary because once the hierarchical system takes control (i.e., once omega decays to a value below 0.5), the hierarchical models that use our modified structured-softmax policy (models 3 and 4) can no longer account for actions that do not conform to one of the highest-level representations of action available to these models. This would result in infinitely poor fits. The errors observed empirically at this late stage of the task were presumably due to lapses in attention and they are not of central interest here, and so we allow for this slight deviation from the model specification to avoid this issue. This was the case for all hybrid models, and so it does not impair comparison between them.

We found that our hybrid version of model 4 provided the best fit to most of our participants (see Figure 5C for fits, and Figure 6 for average predicted behaviour given best-fitting parameters). Of our twelve participants: eleven were fit best by hybrid model 4, and one was fit best by our flat baseline (*flat-history*). Not only did hybrid model 4 provide the best fits to behaviour, but it did so with parameters which we demonstrated to consistently exhibit zero-shot learning. We showed that learning

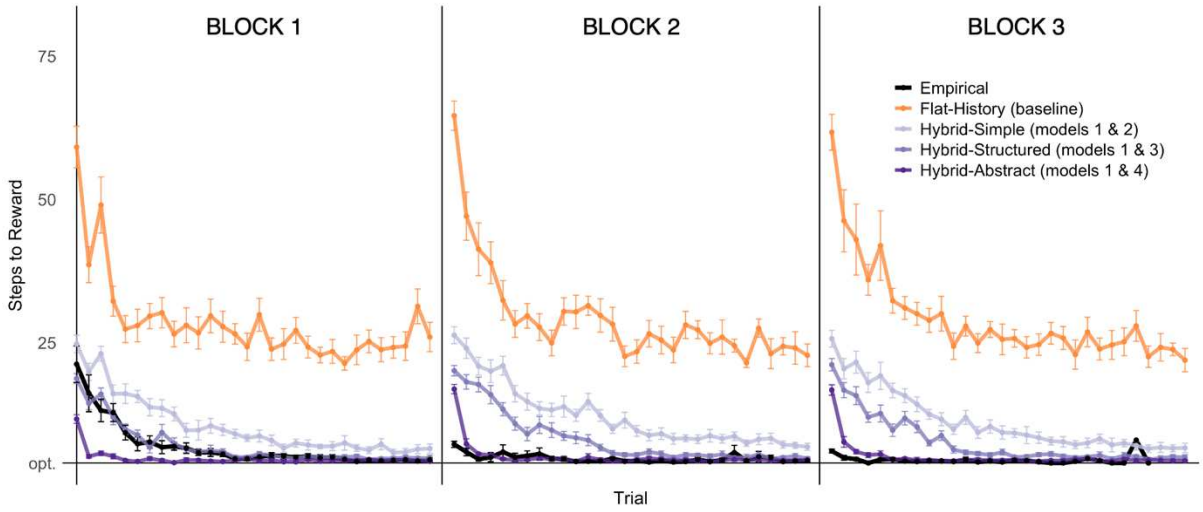

Figure 6 – Average simulated behaviour for each model with best fitting parameters to all 12 participants. For each model, we simulated the behaviour under the best fitting parameters to each participant 20 times, and we averaged over all replications and over all participants. We compare this with the empirically observed average behaviour (in black), which is taken over all 12 participants in both contexts (spatial and procedural).

rate was the primary factor determining the success of model 4 in capturing zero-shot learning. Further, zero-shot learning was best captured by high learning rates (see Figure 5B). The best fitting learning rates of the hybrid model to our 12 participants were close to 1 (mean = 0.97, SD = 0.06), and these learning rates did not deviate significantly from the learning rate which we found to produce zero-shot learning most consistently in model 4 (no significant deviation from the optimal learning rate of 1, $t(11) = -1.74, p = .110$). Thus, the hybrid model achieved generality while still capturing our key behaviour of interest. Our hybrid model therefore fits well to behaviour, and it does so by fitting parameters that we have demonstrated to facilitate immediate generalisation of learned patterns of behaviour to generate completely novel sequences of action.

Hybridised versions of models 2 and 3 performed poorly, and were outperformed by the flat model for all but one participant. This reflects their inability to capture zero-shot learning. Given that these models cannot reliably capture zero-shot learning, not only are individual instances of zero-shot learning unlikely, but all following trials that perform the same sequence of actions are unlikely because these models receive no opportunity to unlearn the previous association between sub-goal and goal. For example, if the original association is repeated, model 3 will learn to solve the task by learning the two paths that implement repetition via the two sub-goals. However, consider a participant that experiences a switch to alternate via the right sub-goal, and then learns immediately what to do via the left sub-goal (i.e., a participant that performs zero-shot learning of the left alternate path). Model 3 would, in this case, be offered no opportunity to learn that repeating via the right is no longer rewarding, and given that model 3 learns only by experience with its environment (and it cannot learn by generalising abstract knowledge), it would expect repetition to be more likely than it was in reality because it expects repetition to still lead to reward. This was an unexpected finding: the hierarchical organisation used by models 2 and 3 was detrimental to their fits to behaviour, and this was owed to the inflexibility of these hierarchies and the omission of the *abstraction* step we outlined in Figure 1.

Discussion

Humans readily learn and produce action sequences based on high-level relational features that cannot easily be accounted for by simple chaining or flat reinforcement learning models. Here, we presented a novel and purely behavioural marker of the otherwise latent hierarchical structure of behaviour; we found that human participants were able to apply learned structural knowledge to generate completely novel sequences of behaviour that met the demands of an evolving environment. This ability to learn to produce novel sequences of behaviour without practice was only captured by a (1) hierarchical reinforcement learning model that contained high-level and (2) relational representations of action, similar to those observed in primate prefrontal cortex (Shima et al., 2007), as well as an (3) ability to abstract learning over multiple states and a (4) preference to explore at high levels of representation. Simpler models lacking hierarchical structure could not capture this aspect of performance, nor could hierarchical models that lacked relational representations of action; all four components listed were necessary. Further, we found that immediate generalisation of the structure of behaviour from one context to another also depended on fast learning rates, and the best fits to behaviour were found by this same model of abstract hierarchy (paired with a flat system to describe initial phases of learning) with near-perfect learning. Learning how to behave in complex and dynamic environments involves progressively building the hierarchies of behaviour necessary to navigate through them, and we suggest that human agents do this not only be *sequencing* lower-level actions towards higher-level representations of order, but also by *abstracting* over actions in order to achieve a flexible, efficient, and adaptive organisation of action.

Hierarchical Organisation, Relational Abstraction

To reproduce the immediate acquisition of novel behaviours, we found that a hierarchical organisation of action was necessary, and we found that this should include representations of the relations between sequence elements and not only simpler chunks of primitive actions. These two components combine insights from the study of motor control in human and non-human primates to provide a more complete view of hierarchical control. Studies investigating the sequencing of action suggest that the brain holds representations for action at several distinct levels of detail (Botvinick, 2007; Koechlin, Ody, & Kouneiher, 2003; Lashley, 1951; Yokoi & Diedrichsen, 2019). For example, representations of individual actions, of chunks of actions, and of sequences of chunks have been found in the motor and premotor areas of the human brain (Yokoi & Diedrichsen, 2019). Separately, *abstraction* has been observed in the shape of relational representations of action that hold information about the relations between the elements of a sequence (such as their position or whether they will be repeated) independently of the actions that make up the sequence found in primate prefrontal cortex (Shima et al., 2007) and in human parahippocampal and cerebellar areas (Kornysheva et al., 2019).

Here, we found relational representations (e.g., repeat vs. alternate) over sequences (e.g., repeat-left vs. repeat-right) composed of chunks (e.g., (NE, NW) vs. (NW, NE)) of primitive actions (e.g., NE, NW, SE, SW). This organisation involves

sequencing of lower-level chunks to establish higher-level representations of order, and abstraction to represent the relations between the lower-level sequences. Evidence has been presented for both of these operations in isolation: (Yokoi & Diedrichsen, 2019) recorded representations of individual movements, chunks of movements, and sequences of chunks (sequencing); and Shima and colleagues (2007) identified individual neurons in primate prefrontal cortex that responded to any sequence containing an alternation between individual, primitive actions (abstraction). Our results imply the use of relational representation of repetition of or alternation between *chunks* of action, which requires sequencing to form the chunk and abstraction to form the relational representation. Our research therefore ties together sequencing and abstraction to demonstrate that both are used in tandem to generate progressively higher-level representations of action and to produce adaptive and flexible hierarchies of behaviour.

Hierarchy and relational structure also form a bridge between the study of sequential motor control (Lashley, 1951) and hierarchical reinforcement learning (Botvinick et al., 2009). To the best of our knowledge, hierarchical reinforcement learning has considered only temporal abstraction (as in the options framework, Sutton et al., 1999) as a method for building higher-level representations of action from lower-level parts. This involves building increasingly high-level representations of action by sequencing together lower-level actions. However, we have demonstrated that relationally abstract representations of action similar to those identified in the brain (Kornysheva et al., 2019; Shima et al., 2007) can and should be included in hierarchical reinforcement learning models to accurately capture human behaviour. Relational abstraction led to a powerful and impressively fast ability to generalise behaviour between contexts in our participants, and it may therefore be of computational benefit for HRL. In particular, relational abstraction appeared essential for the key behavioural target of this paper: zero-shot learning, or the ability to produce entirely novel sequences of action by generalising learned relational representations to new contexts. In this way, a hierarchical organisation led not only to an efficient storage of action that minimised computational cost, but also to a beneficial ability to learn quickly how to adapt that maximised reward earned.

State Abstraction

Abstract representation of action was useful for our HRL models only because of a third component we identified as necessary for immediate acquisition of novel behaviours: state abstraction (Abel, 2019; Andre & Russell, 2002; Botvinick et al., 2009; Radulescu, Niv, & Ballard, 2019). We allowed our most complex HRL model (model 4: abstract hierarchical) to generalise whatever it learned from one context to other relevant contexts. In our task, this meant being able to generalise learning between trials where the sub-goal was on the right and trials where the sub-goal was on the left. Abstraction over behaviour and the generalisation of learning over states are tightly linked. Abstract representations of behaviour are useful because we often want to execute sequences of action that are structurally similar but differ in the low-level details. However, sequences will differ in low-level details only when they are performed in different contexts. Thus, abstraction over behaviour is only useful if we can apply whatever we learn about abstract behaviour to other contexts where it is relevant and useful. For example, I do not need to learn how to brew a coffee anew each time I visit a new kitchen – I can reapply my learnings from one kitchen to another, i.e., I can abstract over states. Further, if the layout of a new kitchen is

different to any I have encountered before, I can still make coffee so long as I represent the order of the high-level steps involved divorced from the low-level actions that would implement those steps (i.e., I hold an abstract representation of the sequence) such that I can adapt the precise low-level actions to match the new layout. To summarise, we argue that abstraction over behaviour and abstraction of learning over states together offer a powerful, adaptive, and efficient framework for learning how to behave. In this study, we show how these two crucial cognitive elements coexist in complex goal-directed action sequences.

Preference for High-Level Exploration

The fourth and final component we identified as necessary for zero-shot learning of novel behaviours was a preference to explore at high levels. In our models, this constraint was a directive, rather than a preference – the two models with our novel structured softmax function were required to select only between the highest-level options available to them in a given state. Exploration at high-levels will generally be more valuable and efficient than low-level exploration. This is most apparent at the extremes: there is little to no value in exploring new methods of reaching out and grasping the handle of a kettle (low-level), but there may well be value in exploring alternative coffee machines or sources of coffee beans. While exploration-exploitation trade-offs are well-established in psychology (Mehlhorn et al., 2015), their interaction with hierarchical representation has not been explicitly considered. In our task, the changes in the environment that prompt exploration are relevant for high-level representations of behaviour, so we cannot disentangle a genuine preference to explore at higher-levels of abstraction from a preference for level-appropriate exploration. Future research could change environments in different ways, prompting a need to explore at distinct levels of abstraction, to clarify this point. However, we see the cognitive efficiency of high-level exploration as a prima facie advantage for a genuine preference for exploration at higher levels. Whether this is correct or not, it seems the case that pruning the action space by exploring at appropriate or high-levels would be beneficial for effectively and efficiently resolving the exploration-exploitation trade-off.

Limitations & Future Directions

Although we identified these four components as necessary for reliably producing zero-shot learning of novel sequences of action, they were not sufficient by themselves to exactly match human behaviour. In fact, our participants showed more frequent zero-shot learning than any of our models. We suggest this limitation arises because our models lack a sophisticated mode of directed exploration that would build on the level-appropriate exploration outlined above. Our participants presumably immediately recognised that the rules of the task had changed upon visiting a goal location they knew to previously hold reward, only to find no reward upon reaching it. They could then rule out the association they previously believed to be true and engage with directed exploration of alternatives, rather than exploring either of the two high-level options they have available to them. This requires incorporating a more sophisticated logic into how our agents explored alternative actions in response to changes in the environment. As already discussed, future research might investigate how more sophisticated exploration interacts with a hierarchical organisation of behaviour to efficiently and adaptively guide choice.

Hierarchical models provided the best account of the key specific target behaviour of zero-shot learning, from the set of models that we compared. However, hierarchical models alone are insufficient to explain all behaviour for the simple fact that in order to form a hierarchy of behaviour one must understand the structure of the environment, and in order to understand the structure of the environment, one must have some experience with it. To resolve this we needed to integrate our hierarchical systems which captured the stable and optimal behaviour observed for the majority of the task with a flat system that could capture the initial phase of learning and any subsequent lapses. In effect, these hybrid models capture the transition people must make from a flat system of behavioural control to a hierarchical one, and our simple arbitration process represents the gestation of the high-level options people come to use. This recalls the “option discover problem” in hierarchical reinforcement learning (Botvinick et al., 2009; Stolle & Precup, 2002), which remains largely unsolved. Although we captured a general transition from memory-based flat control to hierarchical control, we have not explored mechanisms to explain how hierarchies emerge from flat memory-based systems. Recent developments in computational RL describe hierarchical memory systems that divides the past into chunks for efficient recall of goal-relevant events (Lampinen, Chan, Banino, & Hill, 2021). Hierarchical memory suggests a plausible intermediate step between our rather simplistic flat system and our more sophisticated hierarchical agent; memory could be chunked and explored in such a way that associated chunks of behaviour can then be consolidated. Further research is required to investigate how action hierarchies emerge from memory.

Conclusion

To conclude, we present a novel method of measuring the latent hierarchical structure of action from behavioural data alone. Our findings support a novel view of how hierarchies of action are formed in the human brain. Our key result was that people are able to learn completely novel sequences of behaviour with no practice, a process we refer to as zero-shot learning. We combined insights from sequential motor control with hierarchical reinforcement learning to develop a model of goal-directed hierarchical behaviour that could describe zero-shot learning and which showed a number of interesting cognitive properties. First, we demonstrated that a hierarchical organisation itself was necessary for zero-shot learning, as were relational representations of action. This confirms our initial hypothesis that both *sequencing* and *abstraction* were used to build hierarchies of behaviour in the human brain. Second, we demonstrated that abstraction of learning between different contexts goes hand in hand with relational representations of action to allow an efficient, flexible, and adaptive organisation. Third, we showed that adding hierarchical structure to action has important implications for how the exploration-exploitation trade-off is negotiated. In sum, we provided direct behavioural evidence for our proposed latent hierarchical structure of action sequences, and we identified two unexpected additional components that were necessary to explain our behavioural marker of this structure: (1) abstraction of learning between different contexts and (2) level-appropriate exploration. Future research may shed further light on the interactions between hierarchy and exploration, may describe more precisely how we transition from flat memory-based behavioural control to hierarchical control, and may expand further on the benefits of a hierarchical organisation of behaviour that go beyond a mere minimisation of computational cost.

Methods

Subjects

Twelve subjects (mean age = 21.08 years, SD = 2.47; 5 males, 7 females) were recruited to complete both tasks in one sitting. The only inclusion criteria were that subjects were to be aged between 18-35, and of the twelve subjects seven were female, and five were male. The probabilities of observing zero-shot learning by chance under the null hypotheses of no hierarchical organisation/no relational representations were identified by simulation. The chance probabilities were found to be low, ranging from 0.01 (0.03) (Mean/SD) for our flat model to 0.12 (0.11) for our non-abstract hierarchical models. We adopted a highly conservative estimate of 0.5 for zero-shot learning to occur by chance, given that zero-shot learning is ultimately a binary choice between paths and so under a conservative atheoretical view this choice becomes analogous to a coin flip. We performed a power calculation to calculate the sample size required to detect a large effect (Cohen's d=0.8) of zero-shot learning occurrence exceeding this chance estimate, with an alpha level of 0.05 and a beta (power) of 0.8. The large effect and relatively low power here are justified by the functional nature of the test for zero-shot learning; we are testing for capacity, which if present will be highly expressed, and if absent will not. This showed a sample size of 12 participants. Subjects were all told that they would be paid an amount that depended on their performance. In both tasks, performing well meant moving from the starting location to the correct goal in as few moves as possible (the optimum being four). All subjects consented to take part and the study was approved by the relevant ethics committee.

Behavioural Task

Subjects performed both spatial and procedural tasks in one sitting. The order of the tasks was counterbalanced across participants such that six of the twelve would complete the spatial task first, and the other half would complete the procedural task first. Each task began with an initial tutorial section which introduced subjects to the rules of the tasks, how they could navigate around the environments, and the incentive structure.

In the spatial task, subjects could move between rooms by clicking on the door through which they would like to travel. Each room was identical but for a rune which was present in the middle of the room. Each room held a different rune, and these runes were static over all trials such that subjects could learn to place themselves within the map by learning which room was associated with which rune. The objective on each trial was to find a key and to then use that key to open a chest. On each trial, participants were told in which of the two sub-goal rooms they could find the key by providing them with the rune associated with the active sub-goal room. Once they found the key, the queue for its location would disappear from the screen, and the participants would then need to find the chest (i.e., the goal) without any prompt. Once participants found the chest, the trial would end with the delivery of reward. Participants would earn a set number of points for reaching the goal, and they would earn a number of points based on how many doors they

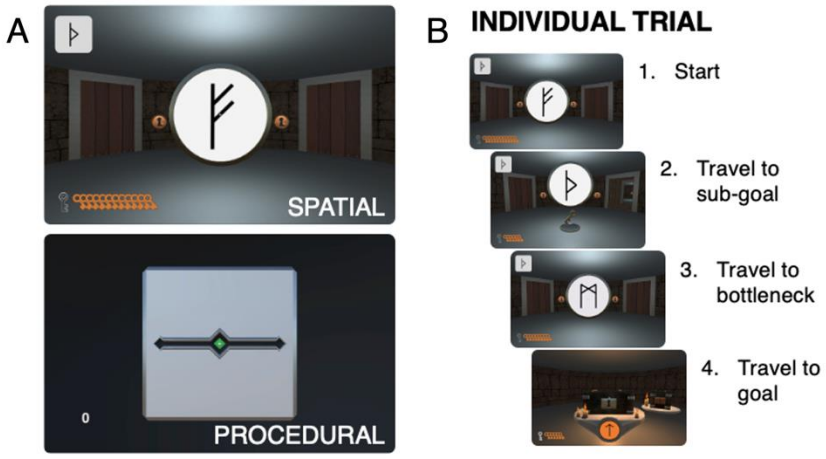

Figure S1 – (A) Initial starting states for spatial and procedural tasks. In the spatial task, participants navigated around a maze of rooms; in the procedural task, participants moved a rod around the faces of a cube. (B) Steps required to complete each individual trial.

opened and travelled through in the environment. If they opened all doors, they would earn no extra points, and they would earn 1 point per door left closed at the end of the trial. This incentivised participants to travel to the goal in as few moves as possible.

In the procedural task, subjects were to move a rod around the faces of a cube by clicking on the edge of the cube-face to which they would like to move. The objective was to move the rod to a cube face of a particular colour (sub-goal) before moving it to a golden cube face (goal). The target sub-goal colour was instructed at the start of the trial. Upon doing so, the trial would end, and points would be earned according to the number of moves taken to complete it. The objective was again to take as few moves as possible.

In both tasks, the location of the sub-goal was randomly allocated on each trial, though it was kept balanced such that there was always an even number of left and right sub-goal trials. From the location of the sub-goal, participants would need to learn to predict where the goal could be found. The sub-goal and goal could be associated in one of two ways: (1) under *repeat*, the sub-goal and goal would be on the same side; and under (2) *alternate*, they would be on different sides. Participants would start with one of these two associations being fixed for the 30 trials that make up the first block. Then, at some point during the first 10 trials of block 2 the association would switch and a fixed 30 trials under the new association would follow, and the same process would happen again on block 3. The order of repeat-alternate-repeat or alternate-repeat-alternate for blocks 1, 2, and 3 was counterbalanced over our twelve participants.

Computational Models

Our flat reinforcement learner followed the algorithm outlined in box 1, and our hierarchical reinforcement learning models all followed the algorithm outline in box 2.

Box 1 – Flat RL model specification

Initialise for all a:

$Q(S, a) = 0$

$S_{origin} \leftarrow S_0$

Repeat until $r = 1$:

$A \leftarrow \text{softmax}\big(Q(S_t)\big)$

$S_{t+1} \leftarrow T\big(S_t, O_{\text{primitive}}\big)$

$r \leftarrow R(S_{t+1})$

$Q\big(S_t, S_{origin}, A\big) \leftarrow Q\big(S_t, S_{origin}, A\big) + \alpha \cdot \Big(r + \gamma \cdot \max_a Q\big(S_{t+1}, S_{origin}, a\big) - Q\big(S_t, S_{origin}, A\big)\Big)$

Box 2 – Hierarchical RL model specification

Initialise for all o and all corresponding S_{init}:

$Q(S_{\text{init}}, o) = 0$

Repeat until $r = 1$:

$h \leftarrow 0$

$O(h) \leftarrow \text{structured-softmax}\big(Q(S_t)\big)$

$S_{\text{init}} \leftarrow S_t$

$r \leftarrow 0$

while S_t is not $S_{\text{term}, O(h)}$:

while $O(h)$ is not primitive:

$h \leftarrow h + 1$

$O(h) \leftarrow \pi_{O_{h-1}}(S_t)$

$S_{t+1} \leftarrow T\big(S_t, O_{\text{primitive}}\big)$

$r \leftarrow R(S_{t+1})$

$Q\big(S_{\text{init}}, O(0)\big) \leftarrow Q\big(S_{\text{init}}, O(0)\big) + \alpha \cdot \Big(r + \gamma \cdot \max_O Q(S_{t+1}, O) - Q\big(S_{\text{init}}, O(0)\big)\Big)$

Model Simulations

To simulate the behaviour of our four models, we established a grid of parameter values for all learning rates in [0.2, 0.4, 0.6, 0.8, 1.0] and all temperatures in [0.2, 0.4, 0.6, 0.8, 1.0]. For each combination of learning rate and temperature, we simulated the behaviour of our four models 20 times. Of these 20 simulated datasets, we then investigated how often zero-shot learning of novel paths following a switch in sub-goal–goal association occurred. The simulations included two blocks of 100 trials, with the sub-goal alternating between right and left every other trial. The switch in association would fall on the first trial of the second block, meaning that

these models had only a single trial to learn the new association before needing to apply any learnings to guide behaviour on the novel post-switch trial.

Model Fitting Procedure

To fit our models to data, we used maximum likelihood estimation. To optimise, we took the negative summed log likelihoods of each individual action given our model, its parameters, and all "experience" up to that action. We minimised this value by adjusting the relevant free parameters for each model using a limited memory BFGS method of parameter estimation (Saputro & Widyaningsih, 2017).

Model Recovery

To ensure that our modelling and fitting procedure was sound and unbiased, we simulated behaviour from our hybrid model given the best fitting parameters for each subject. We then re-used the fitting procedure to fit our hybrid model to these now simulated data to recover the parameters used in the simulation. We repeated this process three. For most participants, we could recover the parameters used to simulate the data with only minor deviations from ground truth (see Figure S2). The only exception was participant 7: the fits for this participant were characterised by a high temperature (beta). Higher temperature means that models explore their environment more, leading to greater noise in the simulated datasets and therefore more noise in the recovery process. With this exception, our simulations accurately recovered the parameters used to simulate data from the hybrid model.

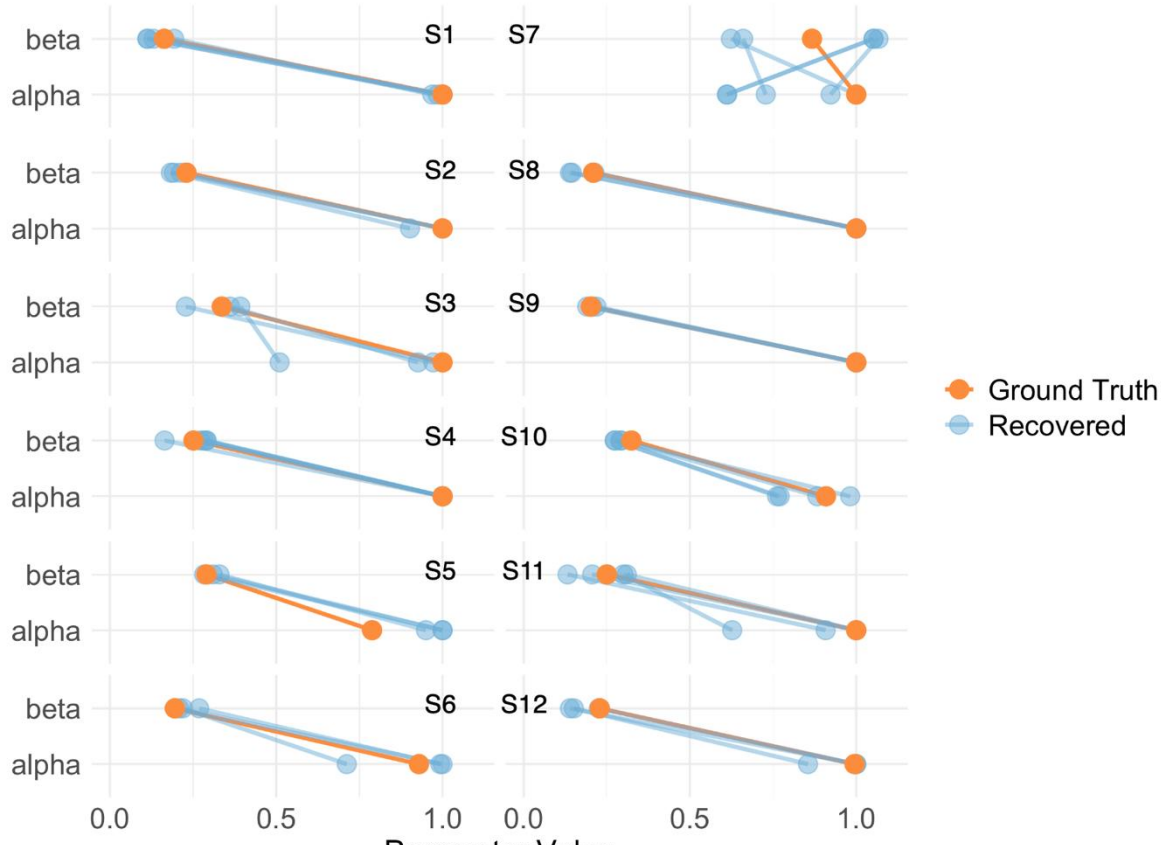

Figure S2 – Ground-truth alongside best-fitting parameters to data simulated from the ground-truth parameters. We simulated three datasets for each subject from our hybrid model with the best-fitting parameters to each subject's empirical data, and then attempted to recover those ground-truth best-fitting parameters three times (corresponding to the three blue lines per participant).

Acknowledgments

We thank Neil Burgess, Jesse Geerts, and Talfan Evans for helpful and insightful conversations during the early stages of this research. The research was funded by a UCL departmental studentship to GW. SP is supported by a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2020-928) and the French National Agency of Research (ANR; FrontCog ANR-17-EURE-0017).

References

Abel, D. (2019). A theory of state abstraction for reinforcement learning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. https://doi.org/10.1609/aaai.v33i01.33019876

Andre, D., & Russell, S. J. (2002). State abstraction for programmable reinforcement learning agents. *Proceedings of the National Conference on Artificial Intelligence*.

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1615–1626. https://doi.org/10.1098/rstb.2007.2056

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Botvinick, M. M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*, 71–77. https://doi.org/10.1016/j.cobeha.2015.08.009

Bullock, D. (2004). Adaptive neural models of queuing and timing in fluent action. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.07.003

Bullock, D., & Rhodes, B. (2003). Competitive queuing for planning and serial performance. *The Handbook of Brain Theory and Neural Networks*.

Cooper, R., & Shallice, T. (2000). CONTENTION SCHEDULING AND THE CONTROL OF ROUTINE ACTIVITIES. *Cognitive Neuropsychology*, *17*(4), 297–338. https://doi.org/10.1080/026432900380427

Diuk, C., Tsai, K., Wallis, J., Botvinick, M. M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(13), 5797–5805. https://doi.org/10.1523/JNEUROSCI.5445-12.2013

760 Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based
761 reinforcement learning. *Current Opinion in Neurobiology*.
762 https://doi.org/10.1016/j.conb.2012.08.003

763 Fuster, J. M. (2008). The Prefrontal Cortex. In *The Prefrontal Cortex*.
764 https://doi.org/10.1016/B978-0-12-373644-4.X0001-1

765 Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The
766 dopamine reward prediction error hypothesis. *Proceedings of the National*
767 *Academy of Sciences of the United States of America*.
768 https://doi.org/10.1073/pnas.1014269108

769 Humphreys, G. W., & Forde, E. M. E. (1998). Disordered action schema and action
770 disorganisation syndrome. *Cognitive Neuropsychology*.

771 Koechlin, E., Ody, C., & Kouneiher, F. (2003). The Architecture of Cognitive Control
772 in the Human Prefrontal Cortex. *Science*.
773 https://doi.org/10.1126/science.1088545

774 Kornysheva, K., Bush, D., Meyer, S. S., Sadnicka, A., Barnes, G., & Burgess, N.
775 (2019). Neural Competitive Queuing of Ordinal Structure Underlies Skilled
776 Sequential Action. *Neuron*, *101*(6), 1166-1180.e3.
777 https://doi.org/10.1016/j.neuron.2019.01.018

778 Lampinen, A. K., Chan, S. C. Y., Banino, A., & Hill, F. (2021). *Towards mental time*
779 *travel: a hierarchical memory for reinforcement learning agents*. Retrieved from
780 http://arxiv.org/abs/2105.14039

781 Lashley, K. (1951). The problem of serial order in behavior. In *Cerebral mechanisms*
782 *in behavior*.

783 Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A.,
784 … Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A
785 synthesis of human and animal literatures. *Decision*.
786 https://doi.org/10.1037/dec0000033

787 Miller, G. A., Galanter, E., & Pribram, K. H. (2017). Plans and the structure of
788 behaviour. In *Systems Research for Behavioral Science: A Sourcebook*.
789 https://doi.org/10.2307/411065

790 Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical*
791 *Psychology*. https://doi.org/10.1016/j.jmp.2008.12.005

792 Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic Reinforcement Learning: The
793 Role of Structure and Attention. *Trends in Cognitive Sciences*.
794 https://doi.org/10.1016/j.tics.2019.01.010

795 Ramkumar, P., Acuna, D. E., Berniker, M., Grafton, S. T., Turner, R. S., & Kording,
796 K. P. (2016). Chunking as the result of an efficiency computation trade-off.
797 *Nature Communications*. https://doi.org/10.1038/ncomms12176

798 Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. A. (2004).
799 Learning and production of movement sequences: Behavioral,
800 neurophysiological, and modeling perspectives. *Human Movement Science*.
801 https://doi.org/10.1016/j.humov.2004.10.008

802 Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y.,
803 & Botvinick, M. M. (2011). A Neural Signature of Hierarchical Reinforcement

Learning. *Neuron*, *71*(2), 370–379. https://doi.org/10.1016/j.neuron.2011.05.042

Rosenbaum, D. A., Kenny, S. B., & Derr, M. A. (1983). Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/0096-1523.9.1.86

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*. https://doi.org/10.1007/s00221-003-1548-8

Saputro, D. R. S., & Widyaningsih, P. (2017). Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). *AIP Conference Proceedings*. https://doi.org/10.1063/1.4995124

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*. https://doi.org/10.1126/science.275.5306.1593

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, *445*(7125), 315–318. https://doi.org/10.1038/nature05470

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003779

Stolle, M., & Precup, D. (2002). Learning options in reinforcement learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/3-540-45622-8_16

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*. https://doi.org/10.1023/A:1022633531479

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/tnn.1998.712192

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. https://doi.org/10.1016/S0004-3702(99)00052-1

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*. https://doi.org/10.1007/bf00992698

Yokoi, A., & Diedrichsen, J. (2019). Neural Organization of Hierarchical Motor Sequence Representations in the Human Neocortex. *Neuron*, 1–13. https://doi.org/10.1016/j.neuron.2019.06.017