University College London

Hierarchical Influences on Human Decision-Making

Author:
Gwydion Williams

Supervisors:
Prof Patrick Haggard
Prof Neil Burgess
Dr Lucie Charles

Declaration:

I, **Gwydion Williams** confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

Abstract

Deciding how to act is complicated because people often hold simultaneous intentions to meet multiple goals. These many goals can be arranged in a hierarchy of goals and sub-goals, and a hierarchy of behaviours can be established to attain them. The hierarchical structure of human behaviour is well established, but the precise form of that hierarchical structure remains unclear. Further, we do not know whether and how this hierarchical organisation of action influences the cognitive processes of deciding between candidate actions. This thesis aims to address these two open questions.

In Chapter 2, I tackle the first of these two questions. Using behavioural experiments in combination with hierarchical reinforcement learning models of behaviour, I demonstrate that people can learn entirely novel sequences of action without practice, and that this ability requires a hierarchical organisation of action built from two distinct operations. First, the brain must sequence low-level components into higher-level routines of action. Second, the brain must have a method of abstracting the relational structure of a sequence away from its content. In sum, this chapter provides evidence for a theoretical framework which can be used to understand hierarchically structured action more deeply.

In Chapters 3 and 4, I tackle the second question: does hierarchical structure influence decision-making? I begin (in Chapter 3) by investigating how hierarchical structure and self-efficacy interact to influence choice between candidate actions. I find that higher level actions are associated with lesser self-efficacy and therefore a lesser willingness to commit to them. This effect arises not only because higher-level actions are more difficult to carry out due to their length, but also because the restrictions that they place on future choices represent a cost. I then (in Chapter 4) investigate whether there are any subjective biases in how outcomes at high or low hierarchical levels are evaluated. I find no overall subjective bias in the evaluation of such outcomes, but I find that social context can prompt strong biases to weight evaluation of outcomes according to their hierarchical level. In sum, I find that hierarchical structure can and does influence decision-making, and I provide evidence for two distinct processes that play a part in this.

These findings establish both a novel theoretical framework for future investigations of hierarchically structured action, and a novel set of interactions between the structure of behaviour and how people make action decisions.

Impact Statement

The way in which the human brain organises behaviour is fundamentally important for understanding how people decide how to behave. This understanding is important not only from a basic scientific point of view, but also for society more broadly. Many of the challenges we now face as a species are at heart a competition between the low-level and immediate outcomes we can enjoy *now*, and the high-level and distant outcomes we must face *later*. The focus of this thesis on understanding the way people structure their behaviour to meet hierarchically arranged goals is therefore important for understanding how we can best meet the many challenges we face day-to-day.

The first part of this thesis outlines and provides evidence for a novel theoretical framework for the study of hierarchically organised action in humans. My results bring together insights from distinct and separate fields of study to provide a more complete view on the way in which progressively higher-level representations of action are formed in the human brain, and I explore the benefits of the organisation used by the brain. The framework outlined in this section of the thesis is of potential use for further investigations of hierarchically organised action.

The second and third experimental sections of this thesis outline several distinct ways in which the organisation of human action influences decision-making processes. These novel findings are relevant not only as the first sign that the structure of human action bears on choice, but also for the design of real-world interventions that aim to aid people in maintaining high-level policies of action. Diverse problems such as addiction, climate change, personal health and exercise might benefit from applying the findings presented here to make people aware of the ways in which their decisions are biased by the structure of their behaviours.

Finally, the experimental paradigms used throughout this thesis provide novel methods of investigating hierarchically structured action. In Chapter 1 in particular, I provide purely behavioural evidence for latent high-level structures, and the approach taken here might prove useful in future investigations along similar lines. The experimental approach taken throughout may prove useful for many fields of psychology and neuroscience given that (as argued in the thesis) all human behaviour follows a hierarchical structure.

In sum, the results and methods presented in this thesis contribute substantially to our understanding of how humans control their behaviour and provide valuable tools for real-world applications and future research.

Acknowledgements

The work in this thesis could not have been completed without the support, help, and guidance of many people. First, I'd like to thank my primary supervisor, Patrick Haggard. Thank you, Patrick, for the many, many interesting discussions, for giving me space and freedom to explore my academic interests and ideas, for your unwavering support, and for your trust. Your mentorship these past few years has been a constant source of motivation. Thank you also to my secondary supervisors, Neil Burgess and Lucie Charles. Neil, thank you for helpful pointers and thought-provoking discussions. And thank you Lucie for introducing me to academic research when I was but a naïve undergraduate.

Thanks to all past and present members of the Action and Body group at the ICN – I couldn't have wished for a better lab to see out these past four years. Thank you in particular to Gaiqing Kong, for being the best of office mates.

A special thank you to the organisers and students of the computational sensorimotor neuroscience summer school, which I attended in August of 2018 before starting on my PhD in the following September. Much of the work in this thesis was shaped and inspired by what I learned in those brief, intense weeks. Thank you, Gunnar Blohm, Paul Schrater, and Konrad Kording.

Thank you to my family: Mum, Dad, Marc a Bethan. Thank you, Mum, for encouraging me to always explore my interests and for your unending support in all things. A diolch i chdi, Dad, am gadarnhau bod gynai addysg da (wrth ddreifio fi dros Gogledd Cymru cyfan!) a pob cyfla posib.

Finally, and most importantly, thank you to my darling Stephi. We can both agree I wouldn't have managed any of this without your love and support. There are too many things to thank you for here, but trust that I appreciate and am grateful for all that you are and all that you have done for me to make this thesis possible.

Table of Contents

Table of Figures

Chapter 1

Introduction

1.1 Overview

Deciding how to act is difficult because we often hold simultaneous intentions to meet multiple goals yet at any given point in time, we can perform only a single action. For example, consider how we decide what to order from a menu in a restaurant. We are unlikely to order more than one dish and deciding what that one dish will be becomes complicated if we consider the wide range of relevant variables (e.g., taste, novelty, price, healthiness, and/or carbon footprint). Conflict between these variables further complicates the decision by requiring that we trade losses in one for gains in another. Classical behavioural economic solutions to decision conflict of this sort involve computing the expected value or utility of each option by taking a sum over the relevant decision variables weighted by preference (von Neumann & Morgenstern, 2007a), and whilst this is indeed a simple and elegant solution, human behaviour is not so simply organised (Brown, Miller, Galanter, & Pribram, 2006; R. P. Cooper & Shallice, 2006; Fuster, 2008; Yokoi & Diedrichsen, 2019). Real-world decisions are not made in a vacuum, and actions are chosen in the context of other policies of action to which we have already committed. For instance, if I decide to become a vegetarian then, for all future decisions of what to eat, all non-vegetarian options can be ruled out. Despite its relevance for behaviour, the sequential and hierarchical nature of human action has received relatively little focus in the science of how we choose to act, and in this thesis, I investigate how the hierarchical organisation of behaviour influences how we choose to behave and how we manage conflict between our intentions.

In this introductory chapter, I first present a review of how behaviour is organised in the human brain as this is centrally important to my investigation of how this organisation influences choice between candidate actions. Lashley's (1951) seminal discussion of the challenges posed by the ubiquity of sequence in behaviour and the failure of reflex chaining accounts of the sequencing of behaviour gave rise to entire fields of research investigating how best to describe and understand ordered behaviour. Here, I discuss Lashley’s contribution by reviewing the pitfalls of flat chaining models of sequential action and the relative successes of hierarchical approaches. I then review hierarchy and abstraction over sequences in greater detail, with a focus on the high-level and abstract representations of action held by the human brain.

Next, I present hierarchical reinforcement learning as a useful normative framework for considering the question of how a hierarchical organisation of behaviour might influence choice. I begin by discussing reinforcement learning and its relevance for the human brain generally, before discussing issues faced by flat (i.e., non-hierarchical) reinforcement learning (namely, the scaling problem) and introducing hierarchy as a solution to these issues.

I will then discuss how insights from investigations of sequential motor control might be integrated with hierarchical reinforcement learning to deepen our understanding of how human behaviour is organised to meet our many intentions. Both fields of research are relevant for any discussion of how human behaviour is organised and controlled, and they are, I argue, mutually informative.

To close this chapter, I discuss how people manage decision conflict in light of the research presented. I begin by setting the problem of conflict between goals that occupy different hierarchical levels more completely, before providing a broad overview of classical investigations of flat and isolated decision-making processes. I conclude this chapter by summarising the primary aims of this thesis, which are to investigate what a hierarchical organisation of behaviour means for decision conflict and how different features of a hierarchy might bear on choice.

1.2 Sequential Motor Control (SMC)

In the early 20th century, the dominant explanation for sequential action came in the form of associative or reflexive chaining models (Ebbinghaus, 2013a; Lashley, 1951), in which each action in a sequence would reflexively trigger its successor. Originating with Ebbinghaus (2013a), the idea of a reflex chain was often tied to language, with Watson (1920) proposing that thoughts were produced by a chain of associations between inaudible movements of the vocal organs. A similar idea was developed more completely by Burtt & Washburn (1918) to propose the peripheral chain theory of language, where language was described as a combination of movements linked together such that any given movement in the sequence acted as a stimulus for what would follow. These and all other chaining accounts share the behaviourist idea that sequence elements serve as conditioning stimuli for successive sequence elements, and whilst these models have lost sway as descriptions of complex behaviour, the fundamental idea remains visible in state-space models of skilled motor control (Buonomano & Laje, 2010; Shenoy, Sahani, & Churchland, 2013b; Sohn, Narain, Meirhaeghe, & Jazayeri, 2019), where the population state of a network of neurons at a given time triggers its state at the next time step according to its connectivity.

Associative chaining models of sequential action suffer from several conceptual problems (Lashley, 1951), and from an inability to explain typical errors in human sequence production (Henson, Norris, Page, & Baddeley, 1996). Conceptually, the primary issue with associative chaining is that these models cannot easily represent different orderings of the same action. Complications arise where the same action appears more than once in a sequence and is followed by different actions with each repetition. Given that associative chaining models require that each action maps deterministically onto the next action in the sequence, repeating the same action but requiring that it triggers different successors is difficult if not impossible. For example, consider how such a system would produce the word “every” – what comes after “e”? One workaround to this issue is to represent each sequence element in terms of its context (Wickelgren, 1969). The two instances of the letter “e” in “every” would, under this scheme, each be encoded along with their neighbours, allowing the first “e” to trigger a “v” and the second an “r”. Whilst this context-sensitive code does solve the issue outlined here, it is computationally inefficient as each action needs to be represented separately for each possible set of neighbours. Further, this context-specific encoding scheme, like simpler context-insensitive associative chaining models, fails to capture typical errors in human sequence production. The most common error in human sequence production is the switching of nearby elements in a sequence (Henson et al., 1996), but given their strictly feedforward nature this is not a natural error for chaining models to produce. The inflexibility of chaining models (Lashley, 1951), and their inability to explain typical lapses in sequential human action (Henson et al., 1996) called for a rethink in how we understand sequence in behaviour.

Lashley (1951) therefore proposed an alternative account derived from observations of the separation of structure from content in language. Lashley observed the ease with which children learn “hog Latin”, which involves transposing the initial sound of each word to the end of the word and adding a long *a* – ke-liay is-thay. Children learn such schemes quickly, and then without hesitation re-structure new words and quickly produce entire sentences that follow this new structure. Lashley also pointed to errors in sequence production, such as Spoonerisms (e.g., “sive drafely”), which demonstrate that entire sequences of words are initiated simultaneously in a way that allows sequence elements to be interchanged. Lashley used these and other examples to argue for a hierarchical organisation of language, which moves up from the vocal movements used in pronouncing a word to the order of words in a sentence to the order of sentences in a paragraph and finally to the order of paragraphs in a discourse. Whilst this was an important step in our understanding of human language, Lashley’s real insight was that this “series of hierarchies” was characteristic of all human behaviour (and, in his own words “almost all other cerebral activity”). There is a syntax of movement that requires explanation, Lashley argued, with the only distinction from language being that in language syntax is more formally defined.

The notion that all human action follows a hierarchical structure has since been further developed. Early responses the idea involved investigating the acquisition of skilled motor control as a process of developing increasingly high-level routines of action (Fitts, 1964; Leonard & Newman, 1964; Schmidt, 1975), which is an ongoing field of research (Yokoi & Diedrichsen, 2019). Others presented *plans* as a high-level realisation of our intentions and considered more deeply how multiple plans are coordinated to produce a single stream of behaviour (Miller, Galanter, & Pribram, 2004). The benefits of hierarchy were also explored: Rosenbaum, Kenny, & Derr (1983) found that hierarchical execution of action facilitated faster and more accurate execution of individual actions; and Ramkumar et al. (2016) observed that monkeys while learning a set of reaching movements would adopt a strategy of performing a sequence of locally optimal trajectories to maximise computational efficiency. Based only on a conceptual understanding of chaining models and on an analysis of sequencing errors, Lashley (1951) correctly inferred the representations and organisation required to produce sequential action in humans. The hierarchical structure proposed has since been further developed and arguments have been made in favour of its informational efficiency (see Bernstein, 1967), leading to hierarchical organisation being the predominant view of how human action is structured.

Central to Lashley's (1951) proposal of hierarchy was a simultaneous and parallel activation of the representations of the actions that make up a sequence. This differs not only from classical chaining models, as discussed, but also from more contemporary proposals of recurrent state-space network models where representations of the elements of a well-learned sequence are never simultaneously active (Rhodes & Bullock, 2002; Shenoy, Sahani, & Churchland, 2013a). Nevertheless, in many cases examined by Lashley, the elements of a sequence were knowable in advance and therefore parallel activation is plausible. The parallel activation of sequence elements is to some extent a necessary feature of any hierarchical system, as activating a high-level representation of a chunk of behaviour means activating a representation of all lower-level parts and the order in which they are to be executed.

One influential implementation of a parallel activation of sequence elements is competitive queuing (Bullock & Rhodes, 2003; Houghton, 1990). Originally developed by Grossberg (1978), competitive queuing (CQ) models are neural network models of sequence production where a competitive choice layer of units is paired with a parallel planning layer of units such that the most highly-active planning unit activates it’s corresponding unit in the choice layer, which then inhibits all other choice units and its own planning unit (see Figure 1-1 – Schematic to outline competitive queuing architecture (taken from Kornysheva et al., 2019). (A) The most active node in the competitive choice layer wins the competition, generates its corresponding action, and is then self-inhibited by an inhibitory connection to its corresponding unit in the planning layer. This allows the next most active nodes to generate the next actions in the sequence, as happens in (B) and (C). This iterative process allows a conversion of the gradient of activations over the planning units into a temporally structured serial output. (D) Averbeck et al. (2002) measured multi-unit activity in prefrontal cortex while monkey drew geometrical shapes, and the results were consistent with the graded parallel preparation of sequential shape segments as described by CQ.). The effect of this architecture is that if a gradient of activations over the planning layer is established such that the ordering of these activities matches the appropriate order of the actions represented by each planning unit and we simply allow the network dynamics to unfold, then the intended sequence will be produced by activations in the choice layer. CQ models of sequential action accurately capture human sequence production, including importantly typical errors in sequential behaviour such as the transposition of nearby actions (Rhodes, Bullock, Verwey, Averbeck, & Page, 2004a).

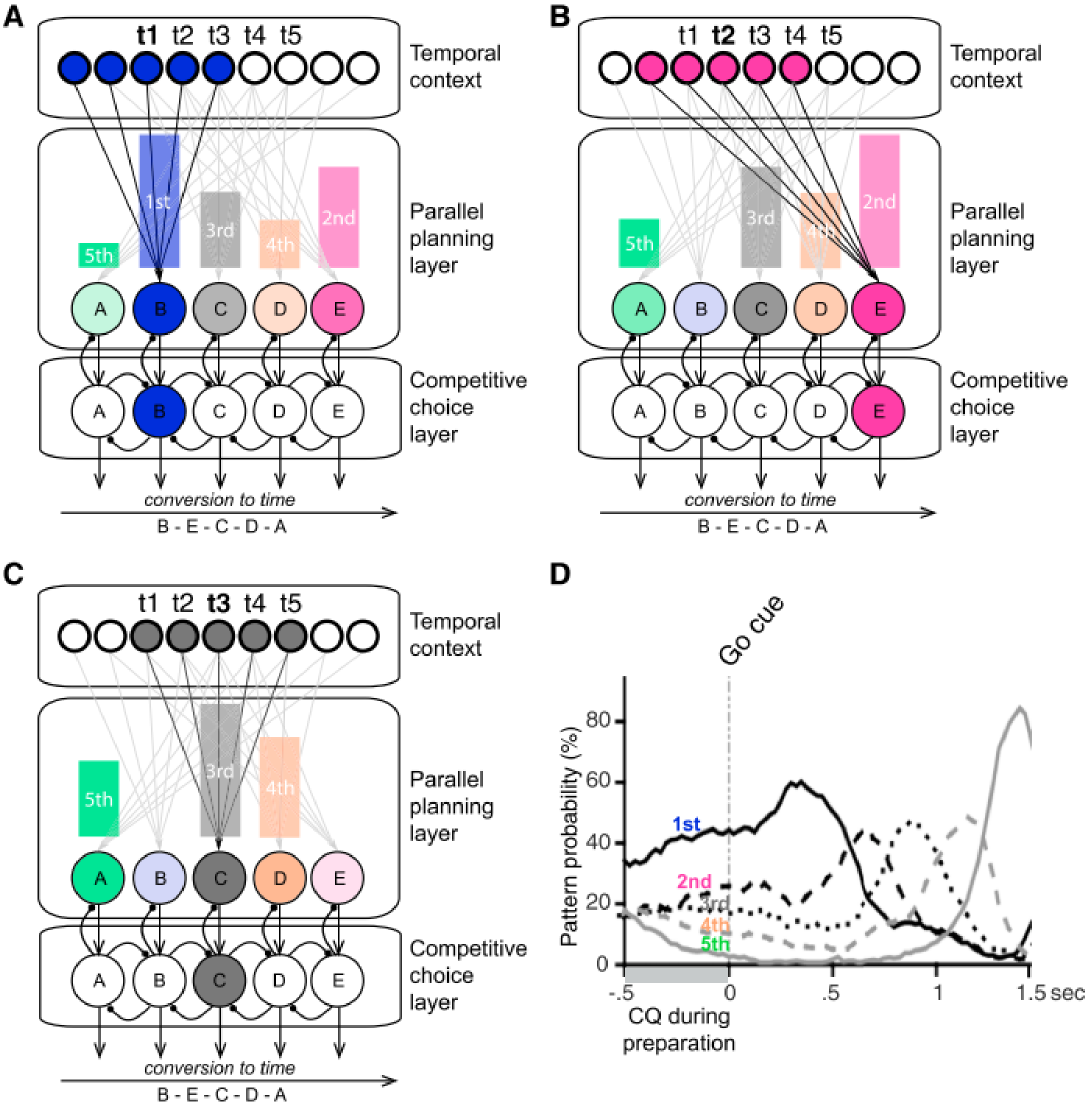


Figure 1-1 – Schematic to outline competitive queuing architecture (taken from Kornysheva et al., 2019). (A) The most active node in the competitive choice layer wins the competition, generates its corresponding action, and is then self-inhibited by an inhibitory connection to its corresponding unit in the planning layer. This allows the next most active nodes to generate the next actions in the sequence, as happens in (B) and (C). This iterative process allows a conversion of the gradient of activations over the planning units into a temporally structured serial output. (D) Averbeck et al. (2002) measured multi-unit activity in prefrontal cortex while monkey drew geometrical shapes, and the results were consistent with the graded parallel preparation of sequential shape segments as described by CQ.

Compelling electrophysiological evidence for CQ was provided by Averbeck, Chafee, Crowe, & Georgopoulos (2002a). They trained monkeys to draw copies of various geometric shapes using routine stroke sequences, such that presentation of a shape cued recall of the sequence of strokes required to copy it. Recordings from the prefrontal cortex in the moments before the monkey started the stroke sequence revealed parallel activations of putative planning units which resembled the activation gradients expected by CQ models. These prefrontal neurons jointly held a representation of the order of the to-be-executed sequence by scaling their activity to match the position of the actions they represent in the sequence. Not only was this parallel representation present prior to sequence initiation, but it evolved in exactly the way an activation gradient in a CQ model would be expected to evolve during sequence execution (with the most active representation being inhibited first, and then the second, and so on). Similar observations have since been made in parahippocampal and cerebellar areas in humans (Kornysheva et al., 2019).

CQ models owe their success to a two-level hierarchy. At a low level, individual actions are represented and executed by individual choice and planning units. At a high level, sequential order is represented in the gradient of activities over all planning units. To demonstrate the importance of this high-level representation of sequence, recall that associative chaining models were unable to deal with repetition (Lashley, 1951). CQ, by contrast, can implement repetition by including a temporal context signal which maps onto the planning units and changes their activity over time (Bullock, 2004; Bullock & Rhodes, 2003). This allows the high-level representation of sequential order implicit in the relative activities of the planning units to evolve with time, and planning units that have already initiated action and have since been deactivated can simply be reactivated by this signal. For another example, recall that switching neighbouring elements is a common error in human sequence production (Henson et al., 1996). This is an easy mistake to make in CQ, as all that is required is that the activations of planning units that represent neighbouring sequence elements (which by design will be close in magnitude) are mis-ordered. Again, this success is attributable to the presence of a high-level representation of order that is divorced from the structure of the network. Changing the order of a sequence is as simple as changing the strength of the activations in the planning units, as this is where our high-level ordinal representation is housed. The success of CQ models of sequential action in capturing human behaviour and explaining brain function is therefore owed to an architecture that makes effective and flexible use of a high-level representation of order.

The study of sequential action in humans has made substantial progress since Lashley's (1951) seminal introduction of hierarchy to the field. We have moved from simple associative chaining models of ordered behaviour towards more successful competitive queuing models of sequential action which not only capture human behaviour (Bullock, 2004; Henson et al., 1996; Rhodes & Bullock, 2002; Rhodes, Bullock, Verwey, Averbeck, & Page, 2004b), but also brain function (Averbeck, Chafee, Crowe, & Georgopoulos, 2002b; Kornysheva et al., 2019). Given this success, it is important to consider what exactly it means for behaviour to be hierarchically organised, and what it means for a representation of action to be high-level.

The most straightforward approach to building a hierarchy of behaviours is to build up higher level sequences from a sequence of lower-level actions. Individual low-level actions can be sequenced together to form higher-level routines of behaviour (Botvinick, 2008; Lashley, 1951; Yokoi & Diedrichsen, 2019), and this rule of high-level representations being formed of lower-level parts is the central architectural principle of any hierarchical system (see *sequencing* in Figure 1-2 –

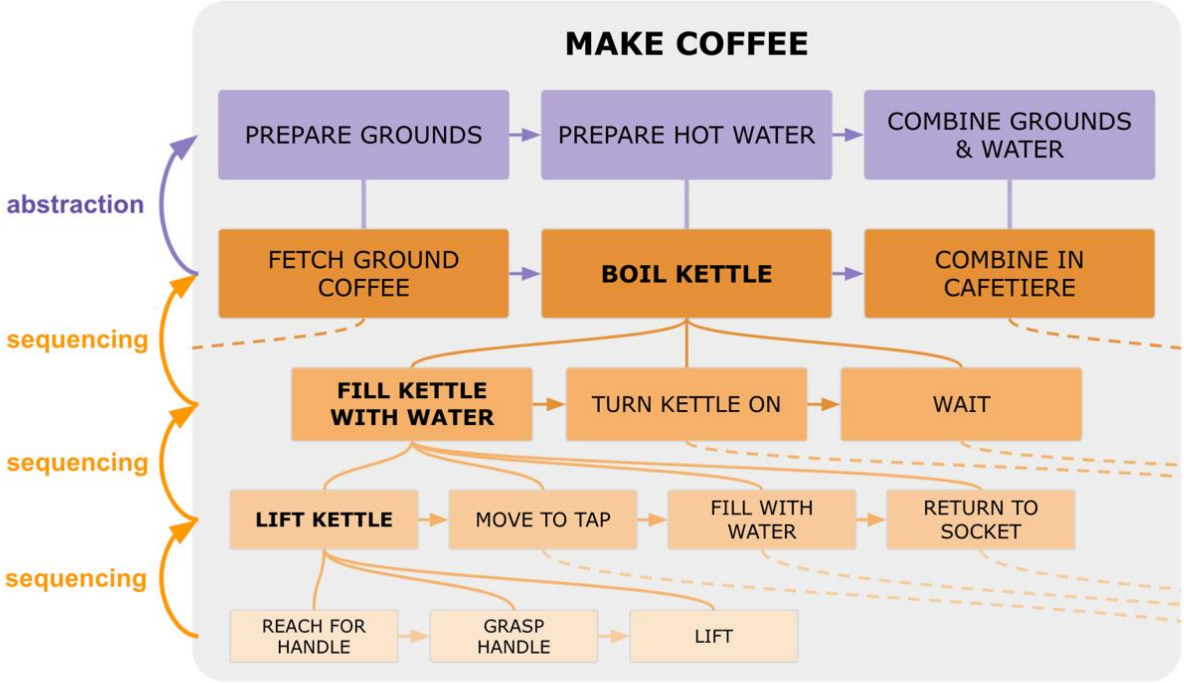


Figure 1-2 – Hierarchy of actions required to make coffee. Higher-level representations of action can come from two distinct operations: (1) sequencing low-level actions (e.g., reach for and grasp the handle of a kettle) can provide higher-level representations (e.g., lift kettle); and (2) abstracting 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).

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).). Higher levels in the hierarchy hold a more compact but less detailed representation of behaviour than lower levels. Higher levels represent broad abstract action thoughts, while more precise motor details are confined to lower levels. Consider the routine of behaviour required to boil a kettle of water. The high-level representation of the sequence of actions required to complete the task (e.g., reach for the handle, grasp the handle, lift the kettle, and so on) need not include the actual motor commands issued to the muscles of the arm to reach for the handle, though it must be able to trigger the circuits that do hold this information in the appropriate order. High-level representations therefore hold no explicit information as to whether a given muscle should contract at any given time, but they do trigger of lower-level actions that will eventually generate this information.

Hierarchies of increasingly high-level representations of action established by sequencing lower level elements have been observed in the human brain. Yokoi and Diedrichsen (2019) trained human participants to perform long sequences of button presses and to then reproduce these sequences in an fMRI scanner. The sequences were organised on three levels; each sequence consisted of four chunks, and each chunk consisted of 2 or 3 finger presses. Here, individual finger presses constitute low-level, primitive actions; chunks represent a higher-level sequence of finger presses; and sequences represent a higher-level sequence of chunks. Using representational fMRI analysis (Diedrichsen & Kriegeskorte, 2017), the authors found separable and distinct neural correlates for individual actions, chunks of actions, and sequences of chunks, providing direct evidence for this style of hierarchical organisation.

Assembling elements into sequences or chunks is clearly one way of producing high-level representations; but is it the *only* way? For a maximally abstract representation of action, one would need to go beyond a sequencing of known lower-level parts and towards more relational representations that abstract away even the individual actions that make up a sequence (see *abstraction* in Figure 1-2 – 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).). For example, if I frequently encounter sequences where I need to alternate between two actions (e.g., right-left-right-left turns while navigating), then it would be computationally efficient to abstract away the actions being alternated in the sequence towards a relational representation of alternation (e.g., rather than right-left-right-left, I would represent A-B-A-B). This *relational* representation holds no information as to the precise actions that make up the sequence, but it does hold information as to how the component actions (whatever they may be) relate to one another. Shima, Isoda, Mushiake, and Tanji (2007) trained two Macaque monkeys to perform 11 different 4-length sequences, each of which followed one of three different temporal structures: (1) “paired” sequences included two pairs of movements (e.g. turn-turn-pull-pull); (2) “alternate” sequences were composed of an alternation between two movements (e.g. turn-pull-turn-pull); and (3) “four-repeat” sequences consisted of a repetition of the same movement four times (e.g. turn-turn-turn-turn). Following a set of trials where these sequences were learned under guidance, the monkeys were then prompted to perform the sequences from memory while recordings were made from neurons in prefrontal cortex. Of 165 task-related prefrontal cells that were recorded, more than half (85) exhibited sensitivity to sequence structure. These neurons fired selectively in the period prior to execution of the first movement of all sequences that followed one of the three sequence structures. That is, some neurons fired selectively for all *paired* sequences, some for all *alternate* sequences, and others for all *four-repeat* sequences. These neurons were firing selectively for a specific relational representation of behaviour, providing evidence for this alternative and abstract mode of hierarchical organisation in the primate brain. The idea here is to encode relations between elements in a way that is content-neutral, analogous to syntax in language.

Whilst the discovery of relational representations of action in primate PFC (Shima et al., 2007) is compelling, translating these findings directly to the human brain is not straightforward given that we are not the same species. Therefore, while we might expect to find similar relational abstractions in the human brain, we should still want direct evidence. Kornysheva et al. (2019) trained human participants to produce specific sequences of finger presses with different orders and timings before reproducing them in a magnetoencephalography (MEG) scanner. Using these recordings in combination with multivariate pattern classification, the authors observed a CQ activation gradient that reflected an abstract template for ordinal position that was used for sequences with different finger presses in different orders. This neural competitive queuing gradient, which was localised to parahippocampal and cerebellar regions, represented the ordinal position of an action within a sequence independently of the action itself. This representation of order is similar to the representations of relational structure observed by Shima et al. (2007), in that both represent the relational structure of a sequence independently of the specific actions that constitute that sequence. These high-level relational representations of sequence allow for a conjunctive coding of behaviour, where abstracted structure can be combined with specific actions to produce distinct instances of a sequence that combine actions in such a way as to adhere to the represented structure. High-level representations of action in the human brain are therefore not formed only by sequencing together lower-level actions. Some representations also abstract away the specific actions that make up a sequence and hold instead only information about the relations between sequence elements (see Figure 1-2 – 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).).

Lashley (1951) in his influential critique of the behaviourist idea that each action in a sequence of behaviour becomes the stimulus for its successor (Ebbinghaus, 2013b; Terrace, 2005) sparked entire fields of research aiming to better understand the role of hierarchy in human action. Introductions of hierarchy to formal models of sequential action enjoyed success in accurately describing human behaviour (Bullock, 2004; Henson et al., 1996; Rhodes & Bullock, 2002; Rhodes et al., 2004b), and even in explaining brain function (Averbeck et al., 2002b; Kornysheva et al., 2019). Further attempts to understand the nature of the high-level representations of action that guide human behaviour lead to evidence for higher- and higher-level representations of action formed by sequencing lower-level parts in the human brain (Yokoi & Diedrichsen, 2019), and to the discovery of relational representations that abstract away the individual actions that make up a sequence towards a content-independent representation of the relations between sequence elements (Kornysheva et al., 2019; Shima et al., 2007). In reality, *all* human action is sequential, and so these successes in understanding how sequences are produced speak to how all human behaviour is organised.

1.3 Reinforcement Learning and Goal-Directed Behaviour

Understanding how behaviour is organised in the brain means understanding the architecture of behaviour and how that architecture is implemented in the brain, but this does not speak to *why* this specific architecture was chosen by evolution nor *what* its effects are on a goal-directed control of behaviour. The process of choosing actions to influence our environments to our subjective benefit is arguably the central reason to have a brain, and therefore understanding completely how people choose between actions to further our goals is vital should we wish to understand it. To investigate the influence a hierarchical organisation of behaviour might have on choice and more generally on intentions to meet goals, we must consider frameworks that can solve the problem of needing to decide between candidate actions at multiple levels towards the pursuit of multiple goals.

In a general sense, what it means to be goal-directed is to maximise subjective benefit by earning reward and avoiding punishment, though what reward and punishment *mean* will depend on the goal. Early descriptions of how people select actions in the face of reward and punishment focussed on the predictive relationships between action and/or events in our environments. Pavlovian conditioning (Yerkes & Morgulis, 1909) involves learning associations between neutral and rewarding or punishing events. Famously, a dog would learn to associate a (neutral) bell with (rewarding) food. While this method of learning would allow the dog to predict the arrival of food, it offers no explicit instruction as to how to act to earn the food. Instrumental conditioning, on the other hand, involves learning how to act to maximise the probability of rewarding events and minimise the probability of punishing events (Skinner, 1935; Washburn & Thorndike, 1912). Under instrumental conditioning, a dog might learn that barking leads to the omission of food but that sitting patiently leads reliably to a treat. The relative values of these two actions would be learned by experience, and the dog could optimise its behaviour by selecting the action that most reliably leads to a subjectively positive outcome. While Pavlovian and instrumental conditioning offer relatively simplistic views on behaviour as being driven deterministically by associations between our environment, our actions, and their outcomes, they do capture two existing classes of conditioned behaviour and they make the fundamental point that people can learn what to expect and how to act from experience with the world.

More recently, computational accounts of conditioned behaviour have drawn heavily from reinforcement learning (RL) models (Sutton & Barto, 1998), which all share the use of a scalar reinforcement learning signal to guide learning. All RL problems comprise (1) a set of world states, (2) a set of actions available to a reinforcement learning agent to navigate these world states, (3) a transition function that defines how actions cause the agent to transition from one state to another, and (4) a reward function that defines where reward lies within the state space (Sutton & Barto, 1998). The objective held by the agent is to discover a policy (that being a mapping from states to actions) that maximises value (that being the expected sum of all future rewards given some policy). RL establishes a normative framework within which to interpret behaviour by (1) making predictions regarding the optimal form of behaviour, (2) suggesting a means by which optimal prediction and action selection could be achieved, and (3) detailing specifically the computations that must take place in service of these two functions (Niv, 2009). Why should we want to adopt a normative perspective on behaviour? There are two primary reasons. First, it is not unreasonable to think that, by evolutionary pressure, the brain adapted to find the optimal solutions to sets of behavioural problems (Kacelnik, 1997). Second, discrepancies between *optimal* and *actual* behaviour can be illuminating as they shed light on the implementational constraints under which people make decisions (Niv, 2009; Tversky & Kahneman, 1974). RL as a normative model thus offers a computational understanding of the optimisation performed by the brain in deciding between candidate actions whilst also describing algorithmically how a principled solution to that problem might take shape (Marr, 1976).

The applications of RL to the brain and behaviour are many in number and generally successful (for a complete review of RL in the brain, see Niv, 2009). Perhaps most famously, in the 90s a connection was made between the dopamine system and RL, inspired by a widely-held belief that dopamine served as the brain's reward signal (Wise, Spindler, Dewit, & Gerber, 1978; Wise, Spindler, & Legault, 1978). In a series of experiments where awake monkeys underwent simple instrumental or Pavlovian conditioning while extracellular recordings of the midbrain were taken, it was found that phasic dopaminergic firing did not only signal the motivational value of rewarding stimuli, but that if these rewarding stimuli were reliably preceded by a predictive stimulus then the dopaminergic response to reward disappeared and was replaced by a phasic burst of dopamine in response to the onset of the predictive stimulus. Further, this shift towards an anticipatory firing pattern in the midbrain was accompanied by anticipatory behaviours such as licking or anticipatory reaching (Ljungberg, Apicella, & Schultz, 1992; Romo & Schultz, 1990; Schultz, Apicella, & Ljungberg, 1993). These data were unified under the *reward prediction error hypothesis of dopamine* (Schultz, Dayan, & Montague, 1997), which argued that the phasic firing of dopaminergic neurons reflects a temporal difference reward prediction error. Indeed, the correspondence is compelling. Temporal difference reward prediction errors occur only when rewarding events are unexpected, just as bursts of dopamine accompanied rewards only early in training where they were unexpected. Similarly, neutral predictive cues should elicit no prediction error until they acquire predictive value, but once they have acquired predictive value the unexpected onset of one of these cues should prompt a prediction error, and thus a burst of dopaminergic activity. These characteristics of dopaminergic activity and their relevance for an RL-like system in the brain have been replicated in several experiments (Bayer & Glimcher, 2005; Hollerman & Schultz, 1998; Takikawa, Kawagoe, & Hikosaka, 2004), and other important elements of RL have been associated directly with the human brain, such as the separable roles of model-free and model-based systems in guiding behaviour (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010). Recent research suggests dopamine also has other roles, such as regulating motor vigour (Da Silva, Tecuapetla, Paixão, & Costa, 2018). RL has succeeded in providing compelling explanations for how the human brain guides behaviour in the pursuit of pleasure.

As ideas from RL become more influential in psychology and neuroscience, it is worth considering how RL research has evolved within computer science (Botvinick, Niv, & Barto, 2009). Here, attention has shifted to focus on the limits of RL and how they might be addressed. One such limit is the *scaling problem*, which describes how basic RL methods do not deal well with increasingly large task domains. As the number of states or the number of candidate actions in a task grow, performance worsens, and for very large state or action spaces tasks become infeasible to solve. In psychology and neuroscience, investigations of the relevance of RL for behaviour have mostly included simple tasks where the number of states and actions is constrained for the sake of tight experimental design. However, in real-world contexts the brain enjoys no such luxury – state and action spaces are vast, and so the scaling problem and the need for a solution must pertain for the human brain just as it does for computational RL.

One influential approach to solving the scaling problem is to make use of *temporal abstraction* (Barto & Mahadevan, 2003; Parr & Russell, 1998; Sutton, Precup, & Singh, 1999a), which supplements the basic RL framework to include temporally abstract actions. These temporally abstract actions represent sets of interrelated lower-level actions (e.g., reach for kettle, grasp handle, lift kettle, and so on), which are cast as a single higher-level routine (e.g., boil the kettle). Under the most popular implementation of temporal abstraction in RL – the options framework (Sutton et al., 1999a) – these temporally abstract actions are named *options*. Options consist of three components: (1) an initiation set, which defines the set of states within which an option can be initiated; (2) a policy, which maps states onto actions whilst the option is active; and (3) a termination condition, which specifies the states within which the option will be terminated and evaluated. For example, an option of making a sandwich would be initiated in states of hunger (the initiation set), would be implemented by the actions necessary to make the sandwich (the policy), and would be terminated when the sandwich was made (the termination condition). Importantly, options can map states not only onto primitive actions, but also onto other options, allowing hierarchies of behaviour to be assembled.

There has been increasing focus in recent years on finding evidence in human behaviour and in the human brain for hierarchical reinforcement learning (HRL). Most success has been found by investigating the nature of hierarchical prediction error signals. Consider what it means to establish a hierarchy of behaviours to meet an end-goal such as “brew a coffee”. To begin, we must decompose the end-goal of brewing a coffee into a series of sub-goals, such as “boil kettle”, “grind coffee”, “mix ground coffee & hot water in cafetiere”, and so on. These sub-goals must be further decomposed into the requisite steps, and this process continues until we reach primitive action. Note that the primary operation here is to decompose one goal into many smaller sub-goals that one must meet in service of the original goal. Note also that although people hold many sub-goals they are defined as such only by virtue of their relation to our original intention to meet the high-level end-goal. More plainly, one does not intend to boil the kettle for its own sake – one does so only as a step in the sequence of brewing a coffee.

There is a subtle but important point here: the sub-goals are not rewarding in themselves, but they are necessary for the receipt of a desired reward at a later date. In some cases, sub-goals can in fact conflict with the actions one would take in pure pursuit of the explicitly rewarding end-goal. For example, consider a delivery driver that must drive south to pick up a package that they must then deliver at a destination that is north of their starting location. Collecting the package moves the driver further away from the location of reward, and towards a location that, while necessary, is not itself rewarding. What, then, is the *motivation* for moving away from reward to collect the package?

The issue of motivation to pursue sub-goals was investigated by Ribas-Fernandes et al. (2011) in a task similar to the delivery described in the previous paragraph. The authors investigated whether the brain makes use of *pseudo-reward*, which is a form of internal or intrinsic reward that motivates the attainment of subgoals, and is therefore distinct from external or primary reward which is available from the environment (Singh, Barto, & Chentanez, 2005). In three neuroimaging studies, neural responses consistent with pseudo-reward prediction errors were observed within the ACC, habenula, amygdala, and NAcc, all of which have previously been implicated in processing temporal difference reward prediction errors (Niv, 2009). These findings confirm that the human brain processes pseudo-reward to motivate the attainment of otherwise unrewarding but necessary sub-goals, which is a centrally important process of HRL for motivating action.

A second prediction made by HRL given a hierarchy of goals and sub-goals is that outcomes relevant to multiple levels of this hierarchy might be observed at the same time, and as a result multiple distinct prediction error signals (which are present at different levels of the hierarchy) may coincide. For example, consider the hierarchy of decisions involved in going out for a meal. At a high-level I must first make a decision between restaurants, and at a low-level once I make it to my chosen restaurant, I must decide between the meals listed on their menu. Once my chosen meal arrives and I start to eat, two simultaneous but distinct predictions are either confirmed or denied: one about the expected quality of the restaurant; and another about the expected quality of this specific meal. A similar setup was used by (Diuk, Tsai, Wallis, Botvinick, & Niv, 2013) to investigate whether simultaneous but separable prediction errors could be measured from relevant regions in the human brain. Thirty participants completed a task where they would first select between two casinos, and then between four slot machines within their chosen casino. If, for example, the participants learned from experience to expect a win from their chosen casino, but they experienced a loss given their choice of slot machine, this one outcome should have produced dissociable prediction errors at each of the two hierarchical levels in the task. As predicted, analysis of activity in the ventral striatum and ventral tegmental area evidenced two simultaneous but distinct and separable prediction errors. These findings paired with those presented by Ribas-Fernandes et al. (2011) clearly demonstrate the relevance of HRL for how the human brain organises behaviour to meet its goals and sub-goals and for how expectations around those goals are adjusted with experience.

HRL is not only an answer to the scaling problem in computational RL, but it also provides a promising normative framework for investigations into how hierarchical behaviour in humans is organised to meet goals and maximise subjective benefit. There is already good reason to think the brain might use HRL-like systems to organise behaviour to meet its goals (Diuk, Tsai, et al., 2013; Ribas-Fernandes et al., 2011), but there are many other aspects of HRL that require further study. Several connections remain to be made, such as identifying the support structures for temporally-abstract actions (or options), and finding evidence for option-specific policies and value functions. Botvinick et al. (2009) suggested that various premotor and prefrontal areas might subserve these functions of HRL: the dorsolateral prefrontal cortex has long been considered involved in guiding temporally integrated and goal-directed behaviour (Hoshi, Shima, & Tanji, 1998; Shallice & Burgess, 1991; Shima et al., 2007; Wood & Grafman, 2003); and the pre-supplementary area and premotor cortex have been shown to carry high-level representations of task set (pre-SMA: Rushworth, Walton, Kennerley, & Bannerman, 2004, and PMC: Muhammad, Wallis, & Miller, 2006). Indeed, research on frontal cortex generally converges on the idea that it serves to represent behaviour at many nested levels of temporal abstraction (Grafman, 2002; Wood & Grafman, 2003), with higher-level representations being housed more anteriorly (Botvinick, 2008; Botvinick, Niv, et al., 2009; Koechlin, Ody, & Kouneiher, 2003), and such a structure would fit well with the architecture of HRL. There is close contact here between the representations and neural structures required by HRL in the brain and those revealed by investigations into sequential motor control in humans, and so in search of a complete description of human behaviour it is worth considering how these two distinct bodies of research might be fruitfully integrated.

1.4 Integration of Sequential Motor Control & HRL

RL and the control of sequential action share in the employment of hierarchy to solve major computational problems. In sequential motor control (SMC), the inclusion of hierarchy (Lashley, 1951) was a necessary step in accurately capturing typical errors in human sequence production (Henson et al., 1996) and in otherwise providing more efficient and flexible accounts of how people produce sequences of action (Bullock, 2004; Rhodes & Bullock, 2002). In RL, introducing hierarchy solved the scaling problem (Botvinick, Niv, et al., 2009) by allowing these models to truncate large state and action spaces by following temporally extended routines of behaviour (Sutton, Precup, & Singh, 1999b). In both cases, the use of temporal abstraction to form higher- and higher-level representations of action has been central to their respective successes. Given the substantial overlap between the systems that these two bodies of research describe, it is worth considering what it would mean to integrate them and whether any new light is shed on how human behaviour is organised and controlled in doing so.

Integrating the study of SMC with HRL as a framework means claiming that the multi-level representations of action observed in the motor and premotor areas (Yokoi & Diedrichsen, 2019) are the neural instantiations of the temporally abstract actions (or options) used in HRL (Sutton et al., 1999b). This is a reasonable claim, as there is a precise match in the form of these two representations: in each case, high-level representations of action map onto a sequence of lower-level actions, leading eventually to primitive action. That is, both frameworks use sequencing (see Figure 1-2 – 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).) to assemble their behavioural hierarchies. That said, there is an important component of HRL that is missed by this connection, that being option-specific policies. Under the options framework, each option holds its own policy of action which maps states onto other, lower-level options in pursuit of a particular goal or sub-goal. For example, the option-specific policy for “collect package” held by a delivery person would map all states onto the actions that move the delivery person closer to the package. Importantly, however, this is not quite the same as an explicit representation of the sequence of actions required to move from a starting location to the package. A policy is a more general representation of the actions one should take within any given state towards the attainment of a given goal or sub-goal, where a sequence is a simpler list of actions to take. Consider what happens if our delivery person mistakenly takes a left turn in a state where the correct action was to turn right. If all they know is the *sequence* of actions required to move to their goal, then any deviation from that sequence leaves them with no information as to what to do next. By contrast, learning a *policy* of action over all states allows the delivery person to learn how to move towards their destination from any state. After a period of learning, the sequence of actions necessary to achieve the sub-goal will be represented in the policy as each successive state will map onto the next sequence element, but a policy of this sort remains more flexible and less deterministic than an explicit representation of the sequence itself, such as the representations of sequence observed in premotor areas (Yokoi & Diedrichsen, 2019). To fully integrate sequential motor control with HRL, the policies described in HRL and the explicit representations of sequence observed in the study of SMC need to be bridged.

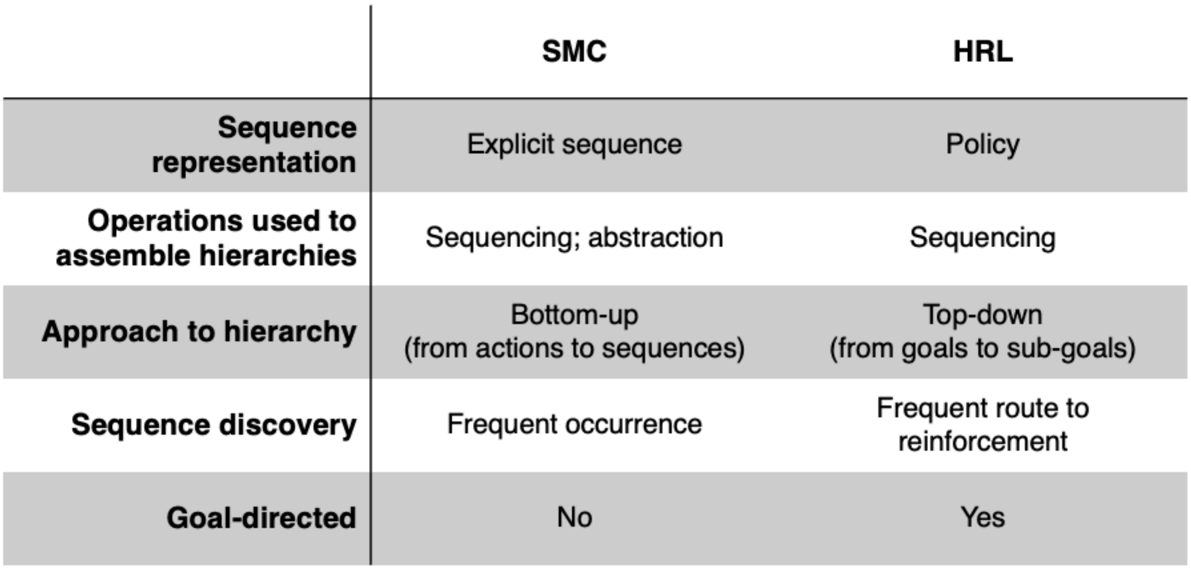
We can reconcile differences between policies of action found in RL and the explicit representations of sequence observed by investigations of SMC if we consider how these representations are formed. In RL, policies are learned and improved by maximising reward and minimising punishment, and hierarchies of action come from a top-down decomposition of a goal into the sub-goals necessary to achieve it. In studies of SMC, however, there are no explicit rewards nor punishments. Instead, higher-level representations of sequence are formed presumably by the frequency with which particular chunks of action are repeated, and hierarchies of action are built bottom-up from this process of identifying repeated chunks of behaviour that might be efficiently grouped up and cast as a higher-level representation. Under this light, these two representations of action and the different modes by which they are built are not necessarily incompatible. It could be the case that the human brain engages both with a top-down process of breaking down abstract and high-level goals into firmer and lower-level sub-goals (as in HRL), whilst also building up a repertoire of multi-level representations of oft-repeated sequences of action (as in SMC). This scheme would fit well with proposed hierarchically arrayed levels of control in the frontal cortex and with a movement from high-level and abstract intention to low-level specific action along its rostro-caudal axis (Badre & D'Esposito, 2007; Koechlin et al., 2003).

This dual process of breaking down goals and building up sequences may be relevant for the option discovery problem (Botvinick, Niv, et al., 2009). This refers to the difficulty of discovering useful options from scratch. Many approaches have been suggested, such as options being innate and genetically specified (Bruner, 1973; Elfwing, Uchibe, Doya, & Christensen, 2007; Wayne Aldridge & Berridge, 1998). Although genetics are likely to play a role, there is also clear evidence that humans and other animals discover useful behavioural subroutines through experience (Conway & Christiansen, 2001; Fischer, 1980; Greenfield, Nelson, & Saltzman, 1972). Most approaches to learning these subroutines from experience involve either an analysis of trajectories that frequently culminate in reward to identify "bottlenecks" in the state space that make for useful sub-goals (e.g., McGovern, 2002), or an analysis of the state space itself, again with the intention of identifying useful "bottleneck" states that make good targets for action (e.g., Machado, Bellemare, & Bowling, 2017; Şimşek, Wolfe, & Barto, 2005). Note that this use of the word "bottleneck" is entirely unrelated to the concept of a capacity-limited stage in a cognitive processing hierarchy, familiar from the attention literature (Broadbent, 1957). These methods that aim to find useful "bottleneck" states break down the state space so that for any given end-goal within that space, a useful set of options are available for achieving it. This maps relatively well onto the top-down process of decomposing a goal into a set of sub-goals. However, relatively little attention has been given to any bottom-up strategy of building higher-level representations from oft-repeated lower-level actions. This second strategy certainly seems to be in use by the human brain (Bullock, 2004; Lashley, 1951; Yokoi & Diedrichsen, 2019), and although it is unlikely to be able to furnish entire hierarchies of action in complex environments, it may efficiently discover and provide useful chunks of behaviour at relatively low-levels. The key difference here between HRL and SMC is that in HRL high-level representations of sequence emerge to the extent that sequences lead to reinforcement, while in SMC they emerge to the extent that they frequently occur.

I have discussed two operations that can be used to form higher-level representations of action from lower-level parts: people can (1) *sequence* lower-level actions so as to produce a higher-level representation of order; or people can (2) *abstract* away the specific actions in a sequence towards a representation of the relations between them (see Figure 1-2 – 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).). Although clear markers of an HRL-like system resembling the first of these operations have been observed in the human brain (Diuk, Tsai, et al., 2013; Ribas-Fernandes et al., 2011), the relational representations of action derived from the second operation do not feature in HRL. Therefore, the human brain may use a framework similar to HRL, which involves both high-level representations of action sequences, and also abstract, content-independent relational representations of behaviour. An intriguing question here is to consider why the human brain would include this additional mode of abstraction, given that the sequential representations favoured by both HRL and SMC both constitute hierarchies *without* abstraction. With more abstraction comes more flexibility and adaptivity to changes in the environment; if I need to produce a completely novel sequence that holds a relational structure already represented in my brain, then a good deal of the challenge involved in producing it is already solved.

HRL and SMC describe similar systems of behaviour, though there are important differences between them. In fact, brain and behavioural data supports the frameworks of both HRL (Diuk, Schapiro, et al., 2013; Ribas-Fernandes et al., 2011) and SMC (e.g., Kornysheva et al., 2019; Lashley, 1951; Yokoi & Diedrichsen, 2019). Therefore, to fully understand how human behaviour is organised and controlled, we must integrate HRL as a framework with findings from investigations into how the brain produces sequences of action. I have outlined two differences between HRL and SMC, centring on the nature of the representations of action held by the human brain, though this is by no means an exhaustive list. Table 1-1 – Summary of differences and relative strengths/weaknesses of sequential motor control (SMC) and hierarchical reinforcement learning (HRL). expands on these differences and highlights the relative strengths and weaknesses of SMC and HRL. Where SMC represents explicit sequences of action that can produce no behaviour other than the specific sequence they define, HRL uses policies of action that can guide behaviour more generally towards a goal with whatever actions are necessary to attain it. However, SMC has been demonstrated to make use of both sequencing (Yokoi & Diedrichsen, 2019) and abstraction (Kornysheva et al., 2019; Shima et al., 2007) in building higher-level representations of action, allowing for a more adaptive organisation of behaviour than is possible in HRL which under the most popular frameworks (Sutton et al., 1999a) uses only sequencing (see Figure 1-2 – Hierarchy of actions required to make coffee. Higher-level representations of action can come

Table 1-1 – Summary of differences and relative strengths/weaknesses of sequential motor control (SMC) and hierarchical reinforcement learning (HRL).



	SMC	HRL
Sequence representation	Explicit sequence	Policy
Operations used to assemble hierarchies	Sequencing; abstraction	Sequencing
Approach to hierarchy	Bottom-up (from actions to sequences)	Top-down (from goals to sub-goals)
Sequence discovery	Frequent occurrence	Frequent route to reinforcement
Goal-directed	No	Yes

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). for a reminder of *sequencing* and *abstraction* operations). More importantly, HRL is goal-directed in a way that the sequences of SMC are not. Thus, only HRL can explain how people *choose* between candidate actions to achieve their goals. Thus, for example, SMC cannot be a theory of volition, while HRL could potentially contribute to a theory of volition. These differences come from the diametrically opposed approaches of these two frameworks to hierarchical organisation – HRL breaks goals down into smaller sub-goals, where SMC build sequences up from lower-level actions. An integration of these two systems would provide a more complete account of human behaviour and would bring together the strengths of each system. This would provide an understanding of goal-directed hierarchical behaviour with the formal rigour and flexible policies of HRL, the efficient relationally abstract representations of SMC, and with both top-down and bottom-up approaches to sequence discovery and organisation included. Surprisingly, however, this integration appears not to have been attempted previously.

1.5 Hierarchical Influences on Choice

In the previous sections, I have outlined how sequential action is implemented by the human brain and have discussed the best accounts we have of the processes involved, which all include hierarchy. I have presented reinforcement learning as a normative framework for the study of goal-directed behaviour in humans, and I have introduced hierarchical reinforcement learning and the options framework as an answer to the scaling problem which applies not only to the computational limits of reinforcement learning as a framework but also to the brain should it use RL-like systems to guide behaviour. Having presented the evidence in favour of a process of building up increasingly abstract representations of action to generate entire sequences, and evidence in favour of HRL systems being present in the brain, I then discussed what it would mean to integrate these two systems. Chapter 2 of this thesis will explore these ideas further, but for the remainder of the thesis the focus is on understanding what the joint system presented and discussed in the previous section means for how people choose how to behave.

The study of decision-making and choice has a long history, but the central challenge has remained the same throughout. Whenever someone needs to decide what to do, they are faced with the challenge of deciding between candidate actions based not on a single decision variable, but on several. In deciding where to live, for example, there are many factors to consider, such as average house price, proximity to friends and family, job prospects, and many more. How can the information from all relevant variables be integrated towards deciding on a single action? Early accounts proposed expected utility as a common currency for the evaluation of all decision variables (Fishburn, 1981; von Neumann & Morgenstern, 2007b). Each variable would be assigned some expected utility, a sum over variables weighted by preference could be taken for each option, and the option of largest summed expected utility would be chosen. Whilst expected utility theory was and remains influential, it assumes that humans act rationally; it assumes that people hold accurate estimates of the value of each variable for each option, that priorities for some variables over others accurately reflect true subjective preferences, that sums are accurately taken, and that values are rationally and sensibly compared. However, human beings demonstrably do not act rationally (Gigerenzer & Gaissmaier, 2011; Hirschman, Kahneman, Slovic, & Tversky, 1983; Kahneman & Tversky, 2019). We make decisions under constraints; we do not hold perfect information about the world, nor do we have the time for an optimal treatment of the information we do hold. This insight led to the identification of heuristics that hasten decision time at the cost of accuracy (Tversky & Kahneman, 1974), and eventually to the development of prospect theory (Kahneman & Tversky, 2018), which explains that we are more sensitive to prospective losses than we are to equivalent gains, as just one example of a perception-like bias in the estimation of value which distorts human decision-making. These irrationalities in our treatment of information whilst suboptimal in the limit of no computational constraints are efficient, fast, and mostly accurate shortcuts to the correct decision (Gigerenzer & Todd, 1999). Nevertheless, they do expose the complexity, intricacy, and subjectivity of our decision processes. However we decide what to do, we do nothing as simple as a rational comparison of expected utilities.

Most research on how people decide has focussed solely on isolated decisions of what to do *now*. Behaviourally reductionist decisions such as deciding whether a cloud of dots are moving predominantly right or left (Gold & Shadlen, 2007) may seem at face value to successfully isolate specific decision processes, but the decision between right or left is made in the context of an experiment and is one element in the sequence of behaviour required to participate in that experiment. As argued in the previous sections, no action is taken in isolation; all actions and all choices are parts of some sequence and slot into a rich hierarchy of behaviours. What influence does the hierarchical nature of action have on how people decide what to do?

By adopting a hierarchical perspective on behaviour, the nature of a decision changes substantially because a single decision is no longer simply a function of present information. Rather, decisions are now a function of both present information and a hierarchy of pre-commitments. Consider the choice a psychology undergraduate makes between pressing right or left in response to a random dot motion stimulus. We name these tasks two-alternative forced choice, but there is in fact nothing forcing the student to choose between only the two alternatives presented to them. They could upturn the table, stand up and walk out, or refuse to engage with the task and wait patiently for the trials to roll by. They do none of these things, because these actions would conflict with higher-level actions initiated by past decisions. At the hierarchical level above this one decision between right or left, we might find a routine of behaviour for complying with task instructions, above which we might find a routine for participating in an experiment, above which we might find a routine for being a good student. The student does indeed decide between right or left because they are adhering to a hierarchy of behaviours that effectively rule out all other alternatives. Progressively lower-level decisions must either adhere to the higher-level actions that sit above them, or those higher-level actions must be changed. These situations that require a hierarchical treatment of behaviour are the primary class of problems of interest to this thesis, and they are too complex for traditional simplistic accounts of choice and behaviour to suffice (Weiss & Shanteau, 2021).

By introducing hierarchy to behaviour, individual decisions are complicated by the fact that they must consider their hierarchical context, but they are simplified by the fact that some options will be ruled out by that context. Consider someone who has decided to become a vegetarian and is perusing a menu. The high-level policy of vegetarianism will rule out all non-vegetarian meals on the menu, which at first glance simplifies the low-level decision of what to eat by shrinking the number of candidates. We can then consider various low-level outcomes such as taste and price towards making a final decision. However, consider what happens if the tastiest and cheapest meal contains meat. The high-level pre-commitment to vegetarianism which was a saving grace in the elimination of half the menu has now become a barrier towards a quick and easy decision to select the most enjoyable meal. Note that the central difficulty here is a conflict between high-level and low-level outcomes. At a high-level, our restaurant-goer may have decided to become vegetarian for the sake of their health, or perhaps the climate, both of which are relatively high-level outcomes as they are realised only over long spans of behaviour in aggregate. At a low-level, taste and price are relatively low-level outcomes which are experienced shortly after a single decision of what to order. Where we have conflict between levels, how do we decide between them?

Resolving conflict between hierarchical levels is not a trivial problem to solve, primarily because of the difference in timescales. High-level outcomes (e.g., climate impact) are by definition further away in time, though they are typically also greater in magnitude than lower-level outcomes which are more immediate but lesser in magnitude. This exchange of time for magnitude of the outcome as we move from low to high levels makes it difficult to see how conflict is resolved, as even by application of temporal discounting (Doyle, 2013; Loewenstein, 1988) no clear answer may be found given that high-level outcomes might enjoy an exponential increase in magnitude to counteract any exponential discounting applied. At heart, this is an issue of resolving competition between the high-level goals we hold and lower-level outcomes that might distract us from attaining them, and this conflict is the central interest of this thesis.

1.6 Operationalising Hierarchy

Hierarchy is a key concept for this thesis. Therefore, a clear experimental operationalisation of hierarchy is required. I take two different approaches to this: I use a (1) *strong operationalisation of hierarchy* to test for one specific proposed hierarchical organisation (in Chapter 2), and I use a (2) *weak operationalisation of hierarchy* to test for effects of hierarchical structure on behaviour without explicitly testing for how the hierarchy is organised (in Chapters 3 & 4). Under the former, strong sense, hierarchy is central to the design of the experiment(s) used, and the experiment aims to isolate specific behaviours that provide direct evidence for a specific hierarchical structure. Under the latter, weak sense, some loosely specified hierarchical structure is assumed to be present, and the experiment aims to reveal how this structure influences action choices. Note that even under a *weak operationalisation*, hierarchical structure is still useful for solving the tasks involved. For example, some rewards in these tasks are contingent upon high-level routines of multiple actions, and so a single reward can only be tied to a set of actions if the entire action sequence is represented in the mind of the agent as a hierarchically organised behaviour. Nevertheless, this strong-weak distinction is important for interpretation of the results, because Chapter 2 uses a *strong* operationalisation of hierarchy to search for support for the specific hierarchical structure I propose. Thus, in the chapters that follow I use this empirical evidence and a clear theoretical framework to investigate the effects of hierarchical structure on decision making without testing explicitly for the presence of that structure.

1.7 Thesis Aims

This thesis aims to investigate whether the hierarchical organisation that the human brain uses to structure action might influence how action decisions are made. More specifically, it aims to understand how people manage conflict between outcomes that sit at different hierarchical levels. In search of a complete answer to this question, I have broken the thesis down into three parts.

First, to verify that the hierarchical organisation of behaviour outlined in the previous sections provides a useful framework for investigating human behaviour, I search for behavioural evidence connecting the fields of SMC and hierarchical reinforcement learning. Such a connection would provide a theoretical framework for analyses of behaviour that I will use in the following chapters. Accordingly, the studies in Chapter 2 show that people do organise behaviour hierarchically, and that relational representations of action can be included in HRL models of human behaviour.

Second, I ask whether features of high-level actions themselves influence how people choose choice between those high-level actions. I present an extension of self-efficacy theory (Bandura, 1974, 1977) to hierarchical behaviour and investigate how perceived capability to carry out a lengthy routine of behaviour influences willingness to commit to that routine. I also consider how precommitment and self-imposed limits on future choice, as defined by the length of a high-level action sequence, influence decisions to commit to that action. The studies in this section show that the length of a policy of action, which scales with its hierarchical level, influences action choices. This influence arises because lengthy sequences are more difficult to perform, and because they limit future choice, which I demonstrate to be aversive.

Third, I ask whether there are any subjective biases to prefer high- vs low-level outcomes. I search for a hierarchical discounting factor, analogous to temporal discounting, but distorting subjective value over hierarchical levels rather than over time. I investigate whether biases in subjective value estimates of outcomes at specific levels of a hierarchy can be elicited by priming, for example by social or attentional contexts. The studies in this section find no evidence to support a general bias favouring high- or low-level outcomes. However, biases in either direction can be prompted by even minimal social cues to attend to specific outcomes.

These three parts amount to first asking how the human brain organises behaviour hierarchically (chapter 2), before asking how perceptions of self-efficacy (chapter 3) and any subjective preferences for outcomes at high or low hierarchical levels (chapter 4) might interact with this framework to influence choice. These two final components address the questions “do I feel able to carry out this action?” and “do I want to pursue this outcome?”. I hypothesise that hierarchical organisation strongly interacts with these issues of choice and value, to determine how people control their behaviour.

Chapter 2

Abstraction and Hierarchy:
How is behaviour organised?

2.1 Introduction

Classically, sequential action has been described as being a process of building up chunks of behaviour by sequencing elementary or primitive actions (Lashley, 1951). When executed, each chunk would activate its motor components in order, and chunks themselves could be sequenced to form progressively higher-level routines of action. This hierarchical structure 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). Recently, distinct representations of individual movements, chunks of movements, and sequences of chunks have been observed in primary motor, premotor, and parietal cortices (Yokoi & Diedrichsen, 2019), supporting not only the utility of the framework but also its use in the human brain. Further, this sequencing of lower-level parts is seen in hierarchical reinforcement learning, which provides a normative framework for investigating such an organisation (Botvinick, Niv, et al., 2009; Botvinick, Weinstein, Solway, & Barto, 2015; Solway et al., 2014; Sutton et al., 1999a). Under this scheme of building high-level routines from low-level parts, the central architectural principle becomes a *sequencing* of lower-level parts (see Figure 2-1 – 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).) and a loss of fine detail towards a more compressed higher-level representation of order (R. Cooper & Shallice, 2000; Fuster, 2008; Humphreys & Forde, 1998; Miller, Galanter, & Pribram, 2017; Yokoi & Diedrichsen, 2019). High-level representations of sequence therefore lack fine temporal detail of the low-level movements involved; they need only store information of the order in which constituent actions must be initiated.

Sequential order is a feature of clear interest for high-level representation, but there is evidence for a more abstract mode of representation at high levels; single neuron and population (neuroimaging) data converge on the notion that, at high levels, the relations between sequence elements are represented independently of the elements themselves (see *abstraction* in Figure 2-1 – 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).). Shima, Isoda, Mushiake, and

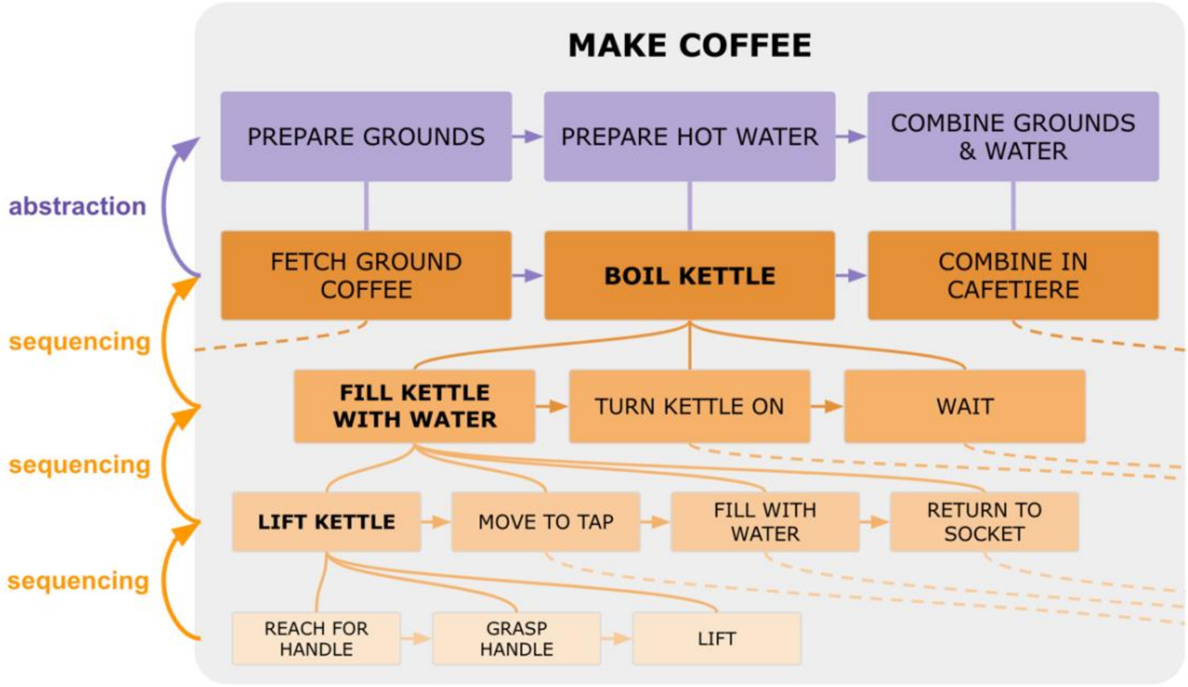


Figure 2-1 – Hierarchy of actions required to make coffee. Higher-level representations of action can come from two distinct operations: (1) sequencing low-level actions (e.g., reach for and grasp the handle of a kettle) can provide higher-level representations (e.g., lift kettle); and (2) abstracting 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).

Tanji (2007) identified prefrontal neurons in the Macaque monkey that fired selectively in the period prior to execution of any sequence that followed one of the three learned sequence structures, and Kornysheva et al. (2019) identified an abstract representation of ordinal position that was used for distinct sequences composed of different actions in different orders. These findings suggest that the brain does not only sequence together lower-level actions towards higher-level representations of sequence, but that it also abstracts over the content of actions sequences towards a representation of the relations between sequence elements.

I propose that *sequencing* and *abstraction* (see Figure 2-1 – 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).) as two methods of building up higher-level routines of behaviour from lower-level actions might be fruitfully combined for a more complete theoretical framework to explain how the human brain arranges behaviour. How might we detect such an organisation from behavioural data? Movement patterns are famously silent about the generative processes that cause 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 I propose a new approach to extracting abstract hierarchical representations from behavioural data based on immediate generalisation of learned sequence structure to produce entirely novel sequences of action to meet completely new challenges, a process I refer to as *zero-shot learning of novel behaviours*. I 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 2-1 – 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).. 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), then 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. I propose this generalisation of structure to produce novel behaviours as a novel behavioural marker of latent hierarchical structure.

In this chapter, I aimed to provide evidence for my proposed framework by searching for zero-shot learning of novel behaviours as a behavioural marker of this latent structure. I report two experiments (experiments 2-1 and 2-2) on goal-directed action which use very different surface presentations, but an identical structure. I observed in both tasks that participants learned new sequence structures from only a single trial, and crucially I found that participants immediately generalised this structural knowledge to produce entirely novel sequences of low-level action on subsequent trials without practice. To verify that this zero-shot learning of novel behaviours was indeed evidence of a hierarchical system that included both sequencing and abstraction, I used computational modelling to explore what were the necessary cognitive components of this learning process. I found that I 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 order, (2) made use of relational high-level representations of action, (3) abstracted learning about these relational representations over multiple states, and (4) directed exploration at appropriate hierarchical levels. In sum, I therefore successfully provide evidence for my theoretical framework that explain how hierarchies of action are assembled, and I find evidence for an influence of the hierarchical organisation I propose on choice in that exploration is targeted at specific hierarchical levels.

2.2 Method

2.2.1 Participants

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. 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 simplest flat model to 0.12 (0.11) for our non-abstract hierarchical models (see Results for more detail on these simulations). 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.

2.2.2 Design & Procedure

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 2-2 – (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.A 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). I used this state map to build two tasks which appeared to be very different (see

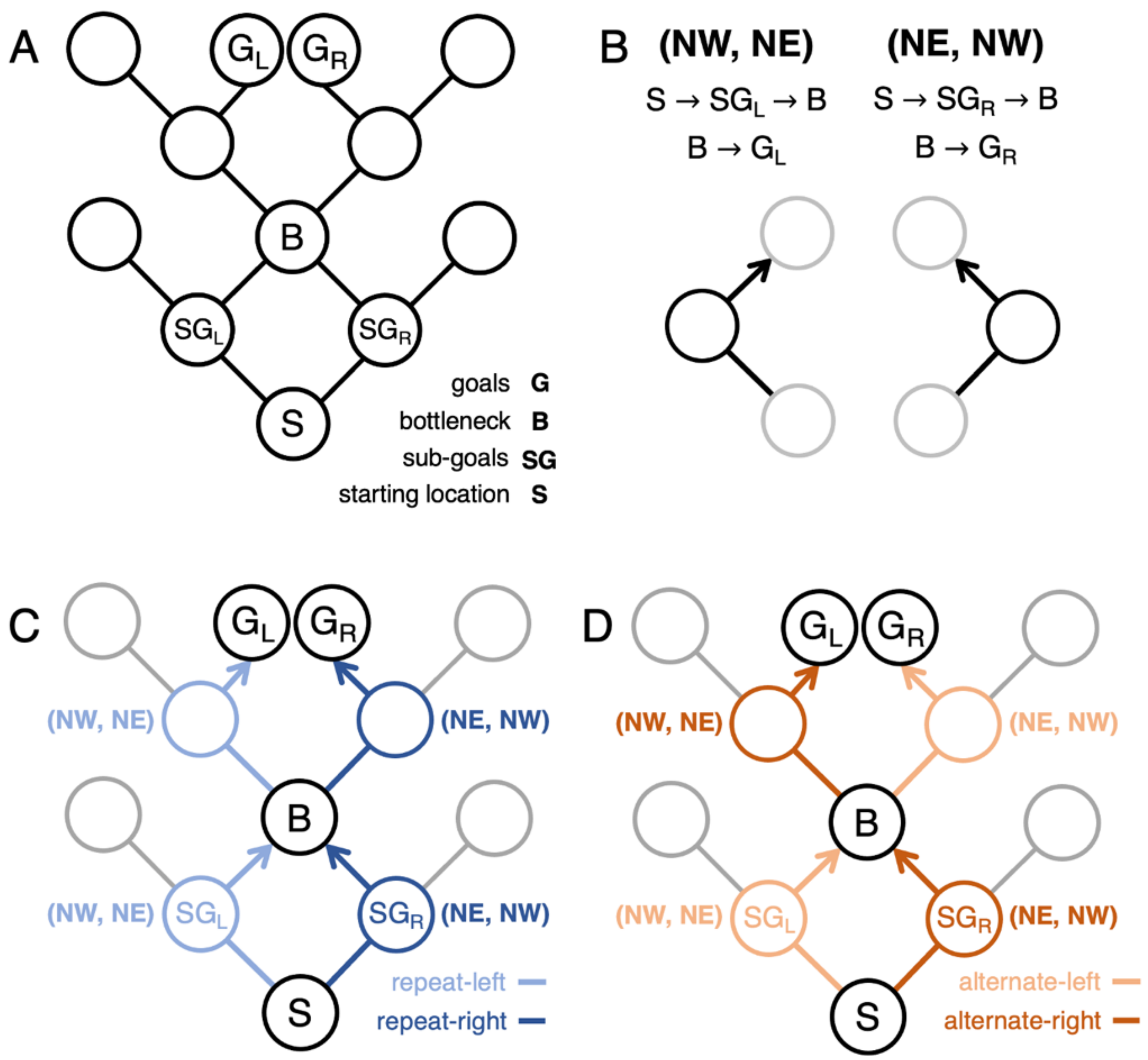
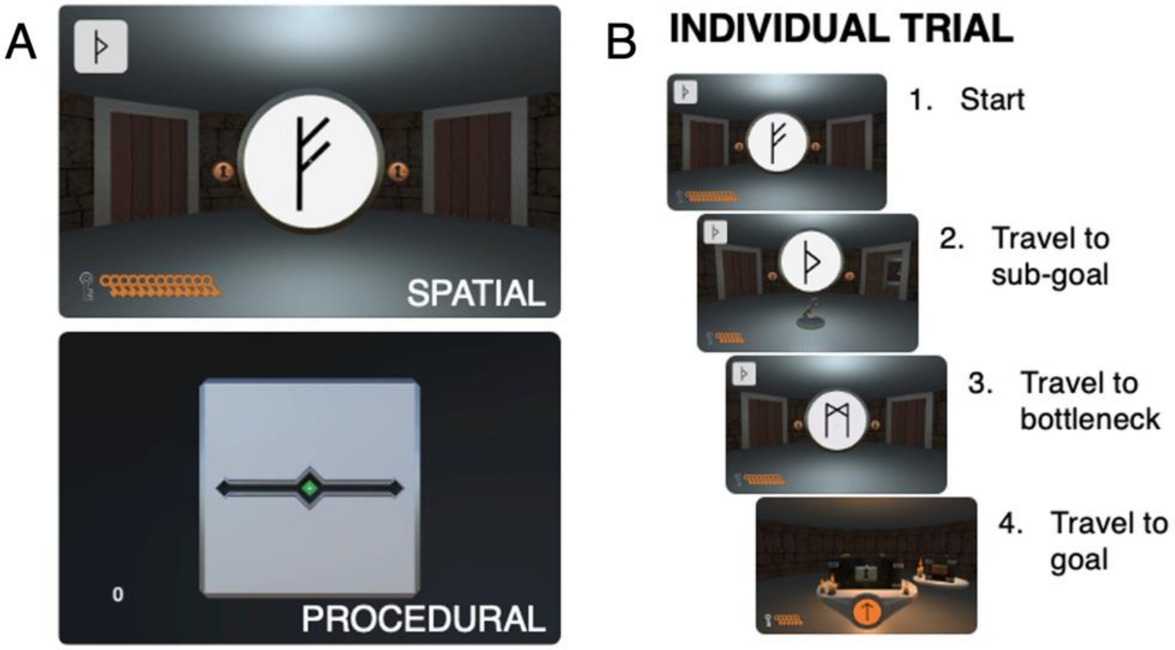


Figure 2-2 – (A) map of state space followed by both spatial and procedural tasks; (B) illustration of useful chunk of actions for navigating to/from the bottleneck state; (C) illustration of the two distinct sequences of action required if the association between SG and G is repeat; (D) illustration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

Figure 2-3 – Surface level appearance for spatial and procedural tasks.) 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 2-4 – 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 2-2). 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. for the full hierarchy). The bottleneck in the centre of the map (see Figure 2-2 – (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.A) 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 2-2 – (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.B). 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 2-4 – 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 2-2). 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.) of my 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 2-4 – 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 2-2). 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.; see Figure 2-2 – (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.B for a demonstration).

Figure 2-3 – Surface level appearance for spatial and procedural tasks.

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 2-2 – (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.B). 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. I 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 2-4 – 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 2-2). 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.): there are two repetition sequences (one each for travelling via the left and right sub-goal, see Figure 2-2 – (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.C), and two alternation sequences (again, one each for travelling via the left and right sub-goal, see Figure 2-2 – (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.D). Finally, there is potential for an abstraction over the *level 2* sequences being repeated in *level 3* such that my participants would represent *repetition* and *alternation* independently of the *level 2* action sequences being repeated or alternated (*level 4* in Figure 2-4 – 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 2-2). 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.). Crucially, participants were never explicitly told whether they should repeat or alternate, but they could derive this information by correctly representing the relation between the sub-goal and the goal, i.e., by representing the hierarchical and relational structure of the task.

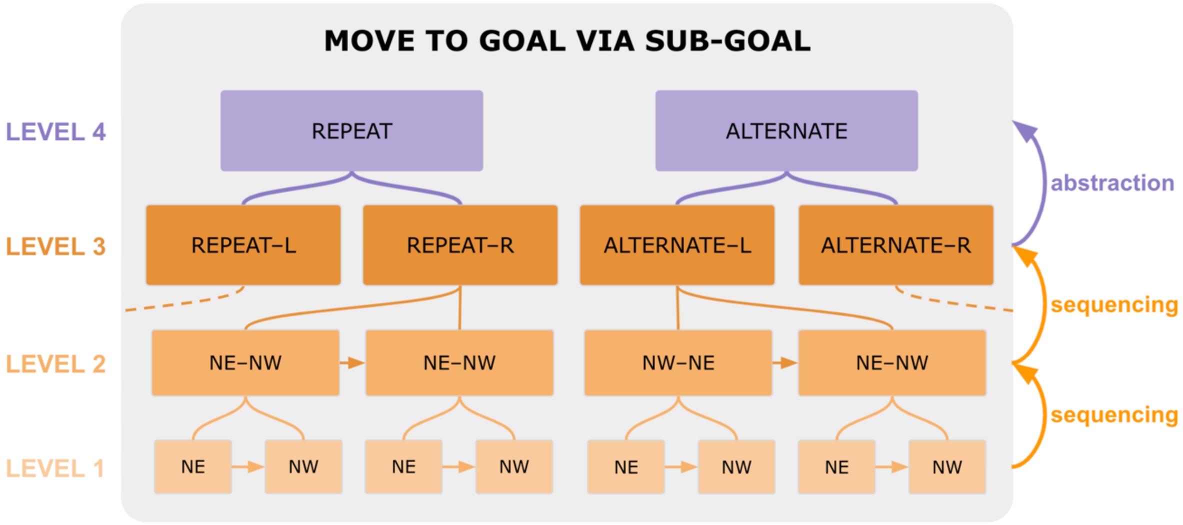


Figure 2-4 – Schematic of the hierarchy of actions targeted by the task design. Level 1 comprises the four primitive actions available in the task. Level 2 contains length-2 sequences that are useful for navigating to/from the central bottleneck (see Figure 2-2 – (A) map of state space followed by both 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.). Level 3 contains

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 2-5 – (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.A). I 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 2-2 – (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.D). I refer to the first trial along this inexperienced path following a switch in sub-goal-to-goal associations the *novel post-switch trial*, and I 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 my dataset the maximum number of trials between a switch trial and its associated novel post-switch trial was four).

2.2.3 Model Simulations

To simulate the behaviour of four models, I 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, I simulated the behaviour of each model 20 times. Of these 20 simulated datasets, I 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.

2.2.4 Model Fitting Procedure

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

2.3 Results

2.3.1 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, IQR = 2.75). Learning was slightly slower on the procedural task (see the shallower rate of learning in Figure 2-5 – (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.B), though behaviour did nevertheless converge on the optimum of four moves to goal. The slower rate of learning on the procedural task 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 2-5 – (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.B), with only minor and infrequent deviations, presumably reflecting lapses in attention.

Our central interest here was in how quickly my participants could recover from a switch in the associations between sub-goal and goal. Specifically, I 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, I was searching for zero-shot learning. This would require a high-level relational representation of alternation and repetition (as in *level 4* of Figure 2-4 – 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 2-2). 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.) that participants could use to adaptively generate completely novel sequences of

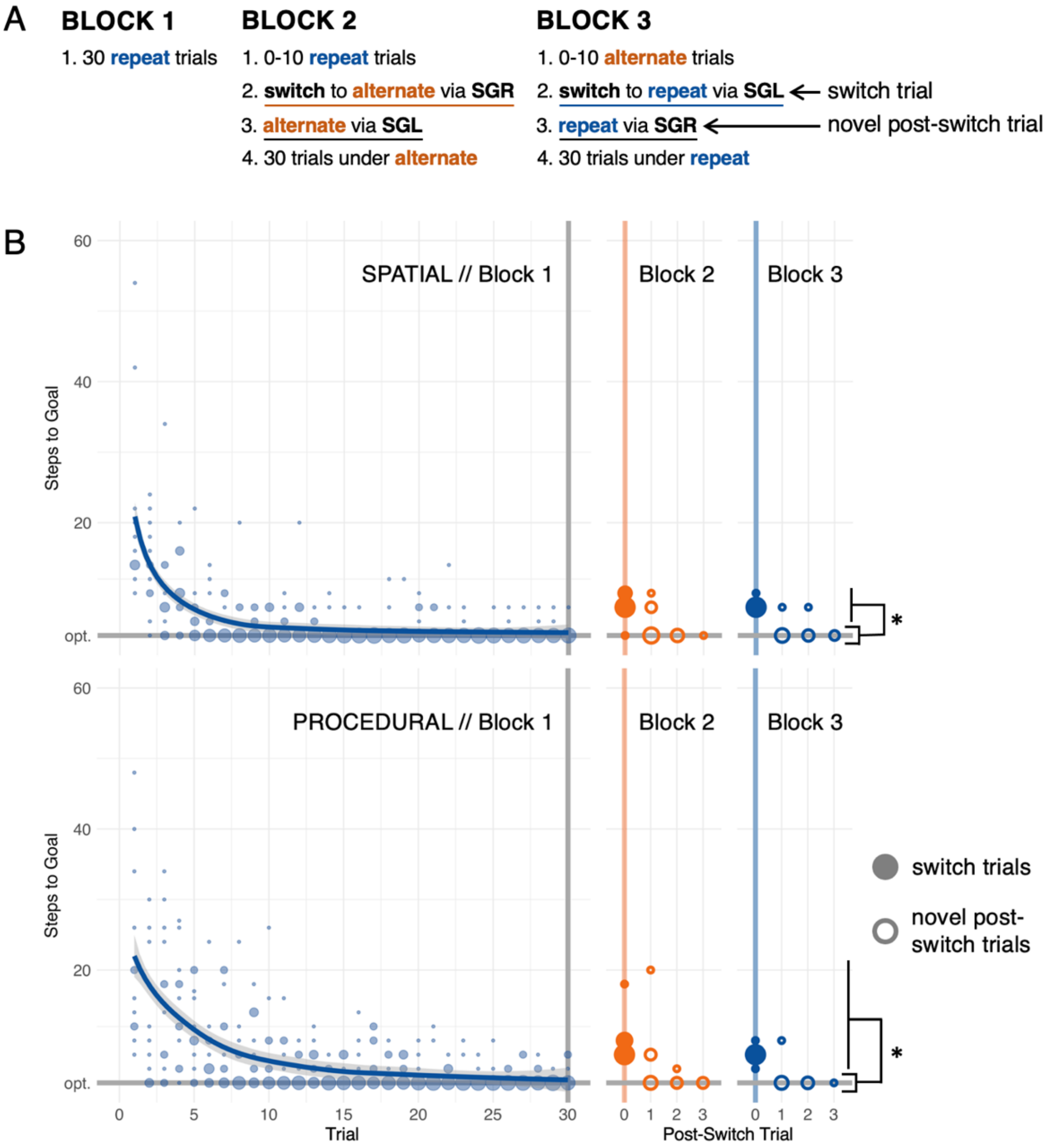


Figure 2-5 – (A) An example of the procedure followed by each of the tasks (note that the order of SG-G associations was counterbalanced over participants); (B) Observed behaviour of 12 subjects on each of the spatial and procedural tasks. The first column plots behaviour of all 12 subjects in the first 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.

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. I found that most of my 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 ($t(11) = 3.22, p = .008$). The number of intervening trials in between the switch and novel post-switch trials 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, $F(1, 46) = 0.94, p = .336$). That is, participants spontaneously generalised learned sub-goal–goal associations to produce entirely novel and optimal sequences of behaviour, an observation I 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.

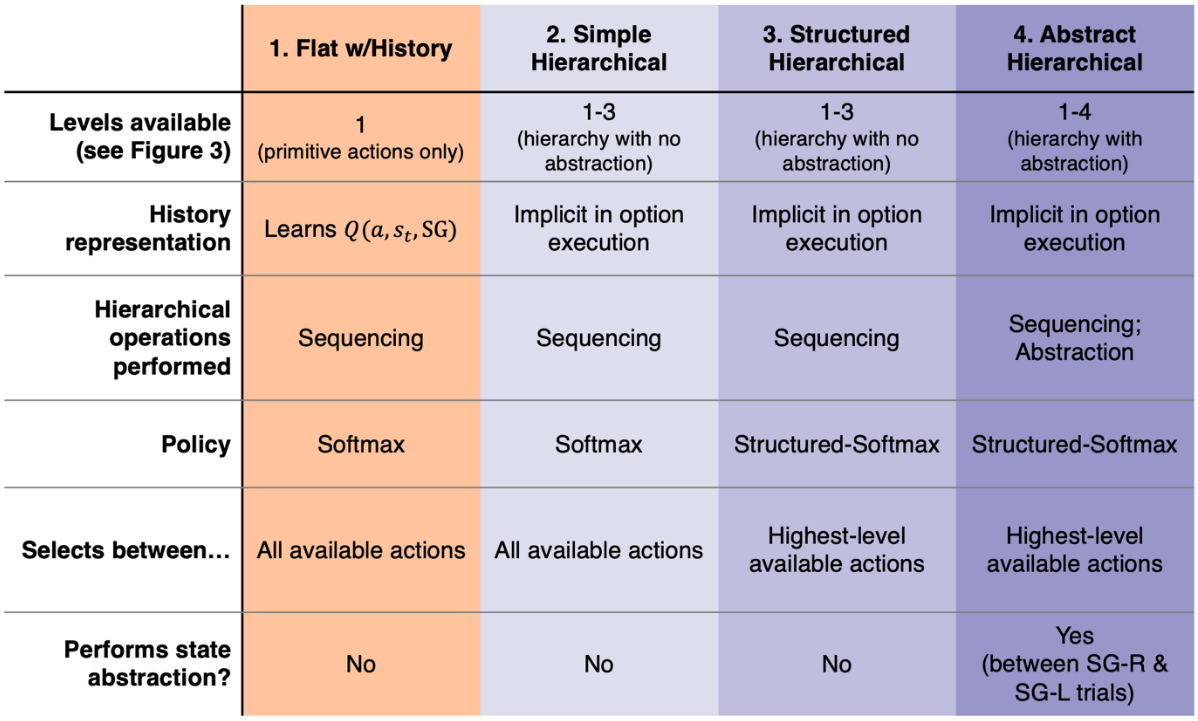
Note that 0.5 is a very conservative chance level for the likelihood of mistakenly performing zero-shot learning of the novel path following a switch. In reality, if my participants understood nothing of the high-level relations between sub-goal and goal, then there would be no reason to think that any 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, I found a mean proportion of 0.77 (SD = 0.29) of novel post-switch trials being optimal over my 12 subjects, providing strong evidence for the ability to perform zero-shot learning of novel sequences of action.

2.3.2 Computational Models

To verify that zero-shot learning of novel behaviours was as I hypothesised a marker of hierarchically organised behaviour that makes use of abstract relational representations of action and to search for any other necessary cognitive components of the process, I built a systematically organised set of four different RL models that aimed to capture my participants' behavioural data (see Table 2-1 – Key differences between my four models. 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 2-4 – 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 2-2). 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.) in the task. 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. I used standard Q-learning over temporal difference prediction errors (Sutton, 1988; Watkins & Dayan, 1992), and modelled participants' choices using a softmax function. Memory was implemented by expanding the Q matrix to include a third dimension of sub-goal location (in addition to two standard dimensions of current state and candidate action). Including this third dimension meant that Q-values for all state-action pairs were sensitive to the location of the sub-goal on any given trial. This first model provided a non-hierarchical baseline against which I could compare the performance of my more complex hierarchical RL models.

The three remaining models were all hierarchical. All three follow the options framework (Sutton et al., 1999a), which supplements the primitive actions available in standard, flat RL with temporally-abstract *options*, corresponding to superordinate chunks of behaviour. The three HRL models had access to particular subsets of the behavioural hierarchy outlined in Figure 2-4 – 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 2-2). 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.. Models 2 (*simple-hierarchical*) and 3 (*structured-hierarchical*) could carry out the options described in the first three levels of the full behavioural hierarchy, and only model 4 (*abstract hierarchical*) held the abstract and relational representations of repetition and alternation present in level 4 of the behavioural hierarchy found in Figure 2-4. Model 4 also abstracted learning over trials where the sub-goal was on the right and trials where the sub-goal was on the left. This was implemented simply by applying all learning updates for each of two origin states (which mapped onto each of the two possible sub-goal

Table 2-1 – Key differences between my 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)

locations) to the other state. That is, if the agent learned that alternating from the right origin state led to reward, it would also learn that alternating from the left origin state led to reward (and vice versa).

I hypothesised that a preference to explore at high rather than low levels was central to the ability to quickly learn and use high-level relational rules, and to test this I 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, my structured-softmax function chooses between only the highest-level options available given the current state of the agent. In practice, models 3 and 4 therefore choose only between the highest-level actions available in a given state.

I aimed to do two things with these four models: first, I used a range of simulations to see how well the various models could reproduce the zero-shot learning of novel paths observed in the behaviour of my participants. Second, I fit 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.

2.3.3 The Necessary Components of Zero-Shot Learning

To estimate how frequently each of my four models could reproduce zero-shot learning by behaving optimally on novel post-switch trials, I simulated the behaviour from each model for a range of parameter values. I 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 I simulated behaviour on the task 20 times. From these simulations, I computed the proportion of novel post-switch trials where behaviour was optimal.

I found that only model 4 (abstract-hierarchical) exhibited proportions of zero-shot learning close to those observed empirically. As expected, my 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 I make no assumptions about the structure of behaviour. Models 2 (simple-

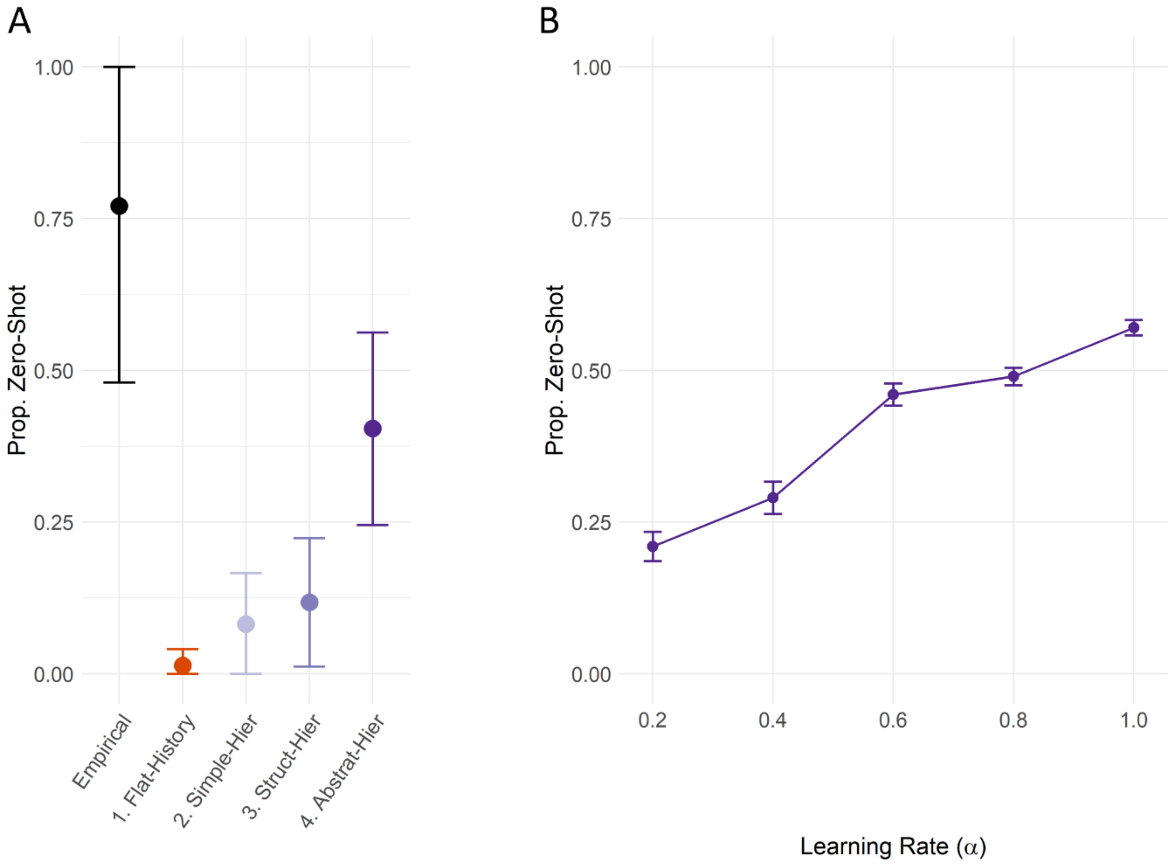


Figure 2-6 – (A) Mean (± SD taken from the proportion of incidences of zero-shot learning over all 20 replications for each combination of learning rate & temperature) proportion of replications that exhibited zero-shot learning for a range of learning rates and temperatures for all four models, with the empirical means plotted for comparison. We see incremental improvements as we increase the complexity of my 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 – I find a monotonic increase in success with learning rate.

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, I 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 2-6 – (A) Mean (± SD taken from the proportion of incidences of zero-shot learning over all 20 replications for each combination of learning rate & temperature) 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 my 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 – I find a monotonic increase in success with learning rate.B). 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.

2.3.4 Model Fits to Complete Span of 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). I therefore additionally fitted these models to the full sequence of behaviour in the task, to investigate their generality, in addition to local fit. However, my hierarchical models learned to use built-in options that were designed to meet the demands of the task, while in reality the brain first needs to learn from experience with the task what these useful options might be. Practically, this means that my hierarchical models are unable to capture the initial period of learning how to solve the task. I 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 then can they build a hierarchical structure able to exploit the relational and structural features of the task. I therefore decided to hybridise my models by using model 1 (*flat-history*) again as a

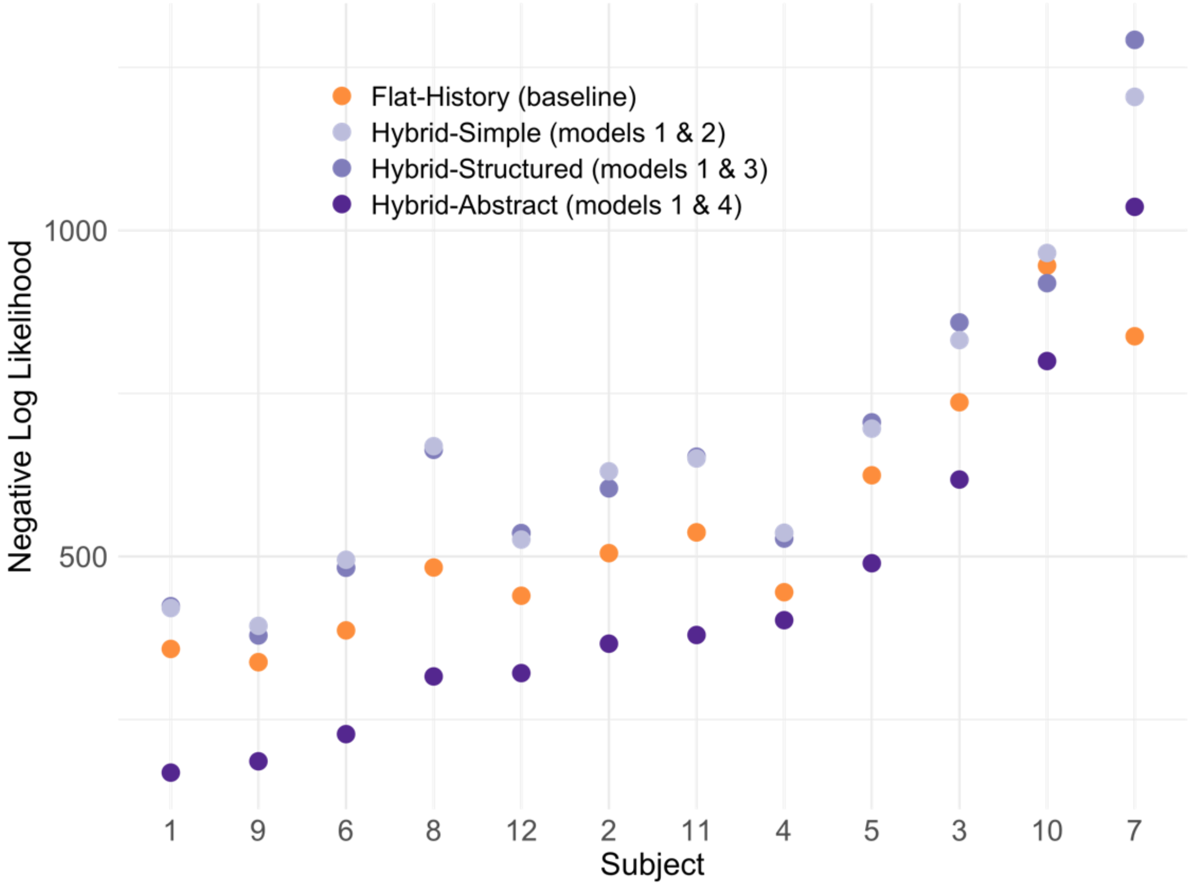


Figure 2-7 – Fits of the flat-history (baseline) and hybrid models to all participants. For 11 of 12 participants, the hybrid-abstract model clearly fits best, with the one remaining participant being fit best by the baseline flat model.

non-hierarchical baseline and to additionally combine each of the hierarchical models in turn with model 1 to establish three hybrid models that would include both flat and hierarchical systems. I included an arbitration process to apportion control of behaviour between flat and hierarchical systems, which was controlled by 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 brain must begin the task with a flat organisation of behaviour (as it does not yet know the structure of the task) but with time discover a useful hierarchy of actions, and this approach of hybridising my hierarchical models was intended to capture this transition from flat to hierarchical behaviour while comparing the ability of each of my hierarchical models to account for global behaviour in the task.

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

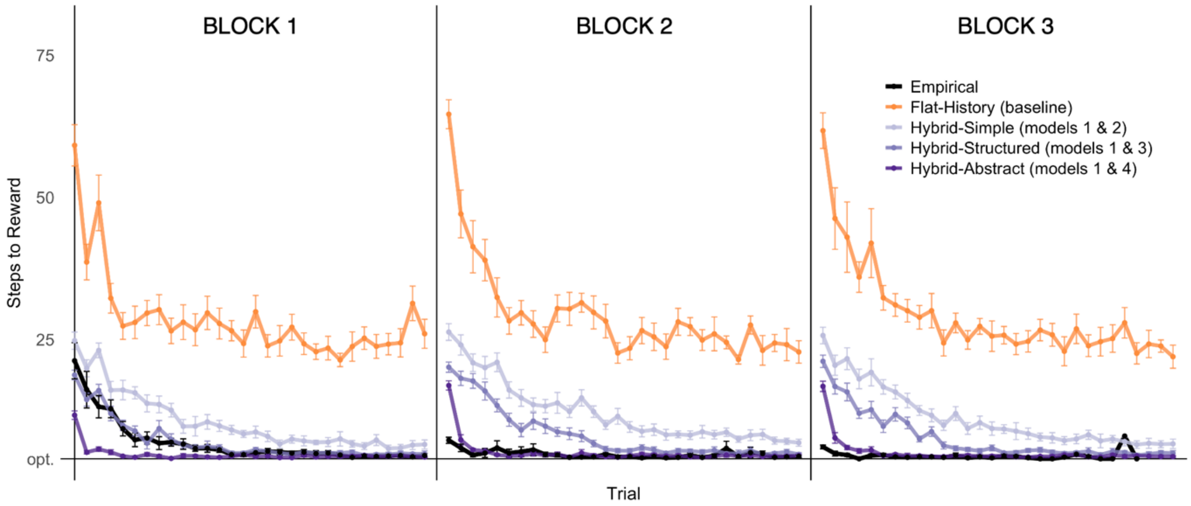


Figure 2-8 – Average simulated behaviour for each model with best fitting parameters to all 12 participants. For each model, behaviour under the best fitting parameters was simulated 20 times, and averages were taken over all replications and over all participants.

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.

I found that the hybrid version of model 4 provided the best fit to most participants (see Figure 2-7 – Fits of the flat-history (baseline) and hybrid models to all participants. For 11 of 12 participants, the hybrid-abstract model clearly fits best, with the one remaining participant being fit best by the baseline flat model. for fits, and Figure 2-8 – Average simulated behaviour for each model with best fitting parameters to all 12 participants. For each model, behaviour under the best fitting parameters was simulated 20 times, and averages were taken over all replications and over all participants. for average predicted behaviour given best-fitting parameters). Of the twelve participants: eleven were fit best by hybrid model 4, and one was fit best by the flat baseline (*flat-history*). In addition to hybrid model 4 capturing behaviour, it also resulted in parameters that consistently exhibit zero-shot learning. I earlier showed that learning rate was the primary factor determining the success of model 4 in capturing zero-shot learning, and zero-shot learning was best captured by fast learning rates (see Figure 2-6 – (A) Mean ($\pm$ SD taken from the proportion of incidences of zero-shot learning over all 20 replications for each combination of learning rate & temperature) 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 my 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 – I find a monotonic increase in success with learning rate.). Consistent with this, the best fitting learning rates of hybrid model 4 to observed behaviour were close to 1 (mean = 0.97, SD = 0.06), and these learning rates did not deviate significantly from the learning rate found to most consistently produce zero-shot 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 the key behaviour of interest. Hybrid model 4 fits well to behaviour, and it does so by fitting parameters that I 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 – for all but one participant, both were outperformed by the flat model. This is owed to 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 repeat, 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 I outlined in Figure 2-1 – 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)..

2.4 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, I presented a novel and purely behavioural marker of the otherwise latent hierarchical structure of behaviour; I 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, I 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 I 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.

2.4.1 Hierarchical Organisation, Relational Abstraction

To reproduce the immediate acquisition of novel behaviours, I found that a hierarchical organisation of action was necessary, and I 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 et al., 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, I 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). My 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. My 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, Niv, et al., 2009). To the best of my 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. I 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. Close contact can be made here to investigations of abstraction over task structures and stimuli within RL problems (Baram, Muller, Nili, Garvert, & Behrens, 2021; Whittington, Muller, Mark, Barry, & Behrens, 2018), though note that I describe abstraction over the agent’s own actions, which is distinct from (but related to) task structure. Relational abstraction over actions led to a powerful and impressively fast ability to generalise behaviour between contexts, 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.

2.4.2 State Abstraction

Abstract representation of action was useful for my HRL models only because of a third component I identified as necessary for immediate acquisition of novel behaviours: state abstraction (Abel, 2019; Andre & Russell, 2002; Botvinick, Niv, et al., 2009; Radulescu, Niv, & Ballard, 2019). I allowed my most complex HRL model (model 4) to generalise whatever it learned from one context to other relevant contexts. In my 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 people often want or need 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 one can apply whatever is learned 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, I 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, I show how these two crucial cognitive elements coexist in complex goal-directed action sequences.

2.4.3 Preference for High-Level Exploration

The fourth and final component I identified as necessary for zero-shot learning of novel behaviours was a preference to explore at high levels. In my models, this constraint was a directive, rather than a preference – the two models with my 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 my task, the changes in the environment that prompt exploration are relevant for high-level representations of behaviour, so I 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, I 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.

2.4.4 Limitations & Future Directions

Although I 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, my participants showed more frequent zero-shot learning than any of my models. I suggest this limitation arises because my models lack a sophisticated mode of directed exploration that would build on the level-appropriate exploration outlined above. My 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 my 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 I 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, I needed to integrate my 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 my 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, Niv, et al., 2009; Stolle & Precup, 2002), which remains largely unsolved. Although I captured a general transition from memory-based flat control to hierarchical control, I 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 my rather simplistic flat system and my 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.

2.4.5 Conclusion

To conclude, I present in this chapter a novel framework for measuring the latent hierarchical structure of action from behavioural data alone, and my findings support my proposed view of how hierarchies of action are formed in the human brain. The key result was that people can learn completely novel sequences of behaviour with no practice, a process I refer to as zero-shot learning. I 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, I demonstrated that a hierarchical organisation itself was necessary, as were relational representations of action, confirming my initial hypothesis that both *sequencing* and *abstraction* were used to build hierarchies of behaviour in the human brain. Second, I demonstrated that abstraction of learning between different contexts goes hand in hand with abstract and relational representations of action to allow an efficient, flexible, and adaptive organisation. Third, I showed that adding hierarchical structure to action has important implications for how the exploration-exploitation trade-off is negotiated. In sum, I provided direct behavioural evidence for the latent hierarchical structure I proposed in Figure 2-1 – 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)., and I identified two unexpected additional components that were necessary to explain my behavioural marker of this structure, one of which outlines a potential influence exerted by hierarchical structure on decisions between candidate courses of action. Future research may shed further light on the interactions between hierarchy and exploration, may describe more precisely how people 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.

2.5 Supplementary Materials

2.5.1 Detailed Procedure

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 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 the twelve participants.

2.5.2 Full Model Specifications

For full model specifications for all flat and hierarchical models, see Box 1 and Box 2 respectively. Parameters for the model specifications are as follows: Parameters are as follows: S_t describes the current state of the agent; h describes the hierarchical level of a given option; o describes an option; O(h) describes the currently active option o at hierarchical level h; π_o describes the policy defined by option o; $O_{\text{primitive}}$ describes the active primitive option; $T(S_t, O_{\text{primitive}})$ describes the state transition that occurs when execution $O_{\text{primitive}}$ from state S_t; α is the learning rate; γ is the temporal discounting rate; O(0) describes the highest level currently active option

Box 2 – Hierarchical RL model specification

```
Initialise for all $o$ and all corresponding $S$ :
    $Q(S, o) = 0$

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:
            $h \leftarrow h + 1$
            $O(h) \leftarrow \pi_{O_{h-1}}(S_t)$
        $S_{t+1} \leftarrow T\big(S_t, O_{\text{primitive}}\big)$
        $r \leftarrow R(S_{t+1})$
    $Q\big(S_{\text{init}}, O(0)\big) \leftarrow Q\big(S_{\text{init}}, O(0)\big) + \alpha \cdot \Big(r + \gamma \cdot \max_{O} Q(S_{t+1}, O) - Q\big(S_{\text{init}}, O(0)\big)\Big)$
```

Box 1 – Flat RL model specification

```
Initialise for all $a$:
  $Q(S, a) = 0$
  $S_{origin} \leftarrow S_0$
Repeat until $r = 1$:
  $A \leftarrow \mathrm{softmax}\big(Q(S_t)\big)$
  $S_{t+1} \leftarrow T(S_t, A)$
  $r \leftarrow R(S_{t+1})$
  $Q\big(S_t, S_{origin}, A\big) \leftarrow Q\big(S_t, S_{origin}, A\big) + \alpha \cdot \Big(r + \gamma \cdot \max_a Q\big(S_{t+1}, S_{origin}, a\big) - Q\big(S_t, S_{origin}, A\big)\Big)$
```

2.5.3 Model Recovery

To ensure that my modelling and fitting procedure was sound and unbiased, I simulated behaviour from my hybrid model given the best fitting parameters for each subject. I then re-used the fitting procedure to fit my hybrid model to these now simulated data to recover the parameters used in the simulation. I repeated this process three. For most participants, I could recover the parameters used to simulate the data with only minor deviations from ground truth (see Figure 2-9 – Ground-truth alongside best-fitting parameters to data simulated from the ground-truth parameters. I simulated three datasets for each subject from my 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).). 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, my simulations accurately recovered the parameters used to simulate data from the hybrid model.

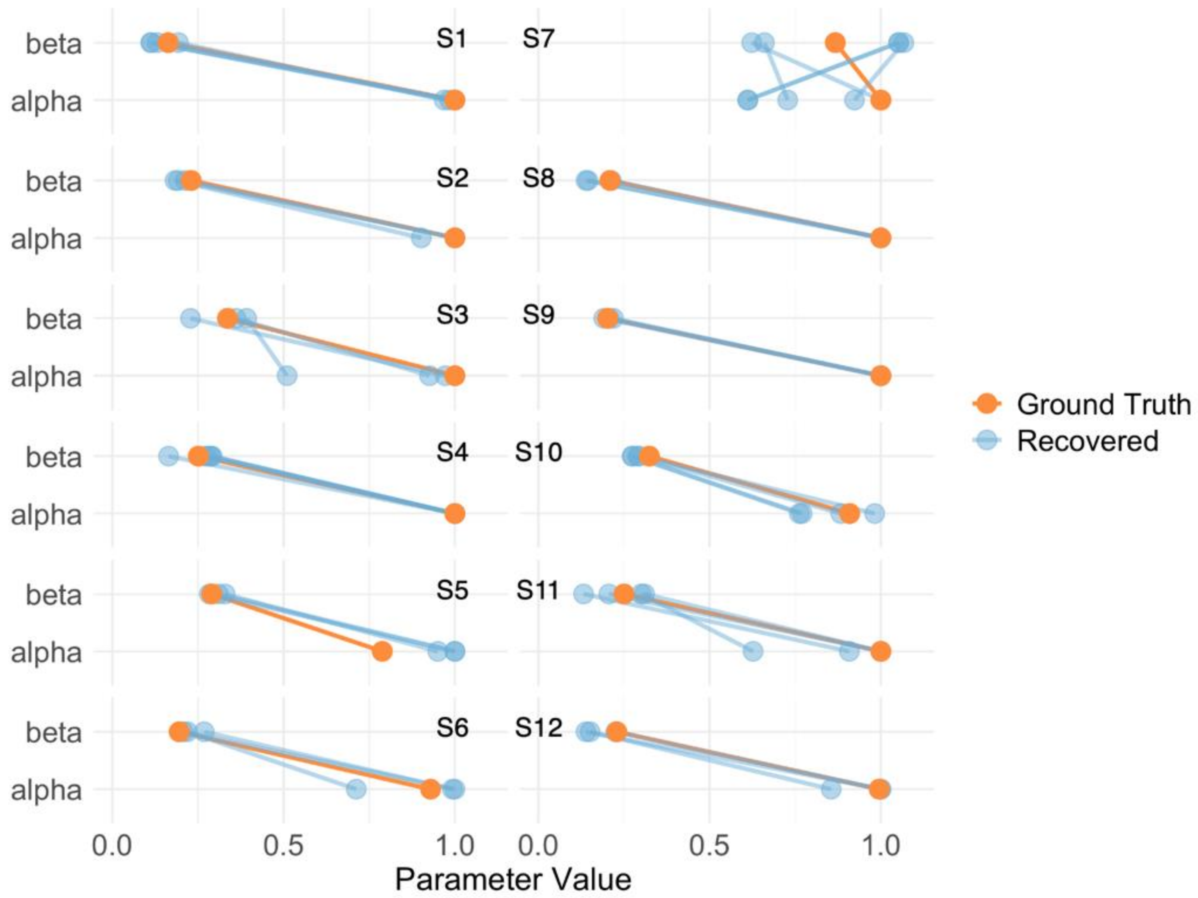


Figure 2-9 – Ground-truth alongside best-fitting parameters to data simulated from the ground-truth parameters. I simulated three datasets for each subject from my hybrid model with the best-fitting parameters to each subject's empirical data, and then attempted to recover those ground-truth best-fitting parameters three times (corresponding to the three blue lines per participant).

Chapter 3

Perceived Capability and Self-Efficacy: Do the features of high-level action influence choice?

3.1 Introduction

Decisions about what to do are not based only on the value of the outcomes that follow from our actions, but also on our ability to carry out those actions. It is no good to decide to compete in the next Olympics because of the high monetary value of a gold medal if I have no real chance of winning. The insight that decisions about how to act are influenced by perceived capability to perform the associated actions led Bandura to develop his theory of self-efficacy (Bandura, 1974, 1977). Self-efficacy judgments are judgments about one’s personal ability to execute contemplated actions (Bandura, 1974, 1977; Kirsch, 1995). Self-efficacy theory accordingly describes the influence these judgments have on behaviour (Bandura, 1984; Holden, 1992; Hyde, Hankins, Deale, & Marteau, 2008; Moritz, Feltz, Fahrbach, & Mack, 2000; Rosenstock, Strecher, & Becker, 1988; Stajkovic & Luthans, 1998). While the link between self-efficacy and behaviour is well-evidenced, it has not been well formalised. Specifically, it remains unclear *how* estimates of capability are derived from the prospect of an action. Further, the hierarchical nature of action has not been considered by, nor integrated with, self-efficacy theory, even though hierarchical organisation presents a range of actions that differ in ways that are relevant for self-efficacy judgments given their hierarchical level. For example, estimating whether I would be able to opt for a vegetarian meal on one specific occasion seems distinctly different to estimating whether I would be able to maintain a policy of vegetarianism in perpetuity.

In Chapter 2, I demonstrated that human action is hierarchically structured, and I provided evidence for a particular hierarchical framework that brought together insights from the study of sequential motor control and hierarchical reinforcement learning. I found that low-level actions are grouped into progressively lengthier, higher-level, and more abstract representations of behaviour. To date, self-efficacy theory has been confined to social psychology approaches and has not made convincing links to computational psychology. For example, self-efficacy theory has not embraced analyses of the different varieties of action, or their internal structure. Rather, the theory tends to homogenise all actions to be described only vaguely as *behaviour*. For example, Moritz and colleagues (2000) conducted a meta-analysis to evaluate the relationship between self-efficacy measures and sport performance.

The precise nature of the tasks varied over the studies included, but all involved self-efficacy judgments followed by a skilled motor task that may or may not have been familiar to the participants beforehand. The self-efficacy judgments made in these tasks were made for distinctly low-level actions; any brief skilled movement is a low-level action with immediate outcomes, and self-efficacy judgments here will be based presumably on the precision of one's motor control. By contrast, Hyde and colleagues (2008) conducted a meta-analysis to estimate the relationship between self-efficacy and prolonged abstinence from addictive substances. Here, the behaviour in focus was distinctly high-level. Sustained abstinence is a long-term and high-level policy of action that will guide very many low-level decisions over the course of an entire lifetime. The outcome is not so immediate, and self-efficacy judgments are presumably made based on the strength of cognitive control, rather than on the precision of motor control. Whether the behaviour in focus is a low-level skilled action or a high-level policy of abstinence, self-efficacy treats each the same and asks simply "could you perform this behaviour?" (Bandura, 1974, 1977, 1984). However, there are important differences between these actions by virtue of their hierarchical level that may be relevant to perceived self-efficacy. More generally, it remains unclear how the hierarchical organisation of behaviour influences perceived self-efficacy. For example, people might feel they can grapple with high-level behaviours, where in fact they cannot, or people might feel they cannot maintain high-level behaviours consistently, where in fact they can.

The principal difference between actions at high vs low hierarchical levels is the length of time over which they span and the number of actions that they therefore prescribe. An individual low-level decision to not smoke a single cigarette controls behaviour for a few minutes and includes a single choice, whereas deciding at a high-level to quit smoking will guide behaviour for an entire lifetime and over very many choices. Intuitively, judging one's ability to carry out each of these two behaviours would be quite different from judging the other. In the former case, one would be estimating one's ability to reject a cigarette a single time, where in the latter, one must estimate one's ability to reject *all* cigarettes *forever*, which may be a less feasible prospect. Despite the clear differences here attributable to the structure of behaviour, applications of self-efficacy theory have mostly ignored this crucial difference. For example, the smoking abstinence self-efficacy questionnaire (Spek et al., 2013), which was derived from Bandura's (1977) self-efficacy theory, asks how likely an individual would be to smoke in a series of 12 distinct situations. While this does capture the generality of a high-level behaviour (i.e., a decision to quit smoking means deciding not to smoke in *all* situations), it does not capture the difficulty inherent in the length of time over which that behaviour spans and the fact that it includes not a single choice, but many. To capture this difficulty, the questionnaire would need to include a final question that asked whether an individual would successfully turn down the cigarettes in all 12 situations listed, and in all other situations they might encounter, without a single lapse. Self-efficacy theory neglects the hierarchical structure of behaviour and casts all behaviour as low-level action, but in so doing it loses the ability to describe how high-level actions influence self-efficacy judgments and therefore choice.

Extending self-efficacy theory to consider hierarchically organised behaviour means describing how perceived ability to execute a single action now is integrated with perceived ability to maintain a long-term and consistent policy of action. One approach is to focus on the here-and-now and to consider the feasibility of individual actions and their consequences (see implementation intention, Gollwitzer, Wieber, Myers, & McCrea, 2010), and to consider high-level self-efficacy as mere repetition of lower-level decisions – as the saying goes, “save the pennies and the pounds will look after themselves”. An alternative approach is to consider the high-level behaviour first, and to consider the feasibility of applying high-level constraints on low-level decisions (for one example, see Reagan’s “Just say no” campaign against drugs, Bourne, 2008). It remains unclear whether self-efficacy is built from the bottom up (by repeatedly dealing with lowest-level action decisions) or from the top down (by having a general constraint at a high hierarchical level that guides each low-level decision), or by some combination.

In this chapter, I extend self-efficacy theory to consider the hierarchical structure of action in order to investigate how the hierarchical level of an action influences self-efficacy judgments. Over two experiments, I investigate two components of high-level action that I hypothesised would contribute to notions of self-efficacy. Both components are derived from increases in the length of a behaviour as we move into higher hierarchical levels. That is, they deal with hierarchy as a simple cumulation of multiple low-level decisions. First, I investigated how the compounding effect of repeating a difficult action multiple times, as prescribed by high-level policies of action, might influence self-efficacy judgments. Second, I investigated whether the restriction of future choice that comes with a commitment to a lengthy high-level policy of action is aversive and thus dissuades high-level commitment. The first of these is a straightforward extension of self-efficacy to include the fact that successfully performing a difficult action multiple times is more difficult than performing it only once. The second idea requires an additional qualitative change to self-efficacy theory, to capture the effect that a commitment *now* has on a decision made *later*. By committing to high-level courses of action (such as quitting smoking), people are in effect restricting all future relevant decisions so that they must conform to that commitment. However, people and environments change, so there are good reasons to be averse to such a restrictive commitment.

I test the hypothesis that these two components (cumulation and constraint) discourage choices favouring high-level policies of action. This is shown in two experiments, both of which present participants with a choice between lengthy high-level policies of action and a sequence of individual low-level actions of equivalent length. I take a novel approach here which is counter to the general approach of measuring self-efficacy by asking for subjective judgments of perceived capability to carry out an individual action (Bandura, 1974, 1977, 2006). Instead, I measure self-efficacy by measuring its influence on choice. I pit two policies of action against one another, and I make one of the two more valuable but more difficult to achieve. This makes a choice to commit to this more demanding but more rewarding policy equivalent to saying that the difference in reward trumps the difference in one’s perceived ability to carry out that policy. That is, I quantify self-efficacy by requiring people to estimate the level of difficulty at which increased reward is no longer worth the risk of carrying out the policy to which they have aspired.

By then adjusting the difficulty level of a risky policy (while keeping rewards constant), it is possible to measure quantitatively how a specific factor of interest influences self-efficacy. In the first of my two experiments, I vary both the length of the candidate policies and the difficulty of the individual actions they prescribe. I find that both of these manipulations predict choice via their influence on self-efficacy. In the second experiment, I make the task environment non-stationary to include an incentive to avoid restricting future choice, and find, as hypothesised, that people will deviate from optimal solutions towards not tying the hands of their future selves.

3.2 Experiment 3-1

3.2.1 Methods

3.2.1.1 Participants

The present study was approved by the UCL Research Ethics Committee, and it was hosted online. Participants were recruited via Prolific (www.prolific.co), and all were then redirected to a personal website where the experiment was hosted. All participants provided informed consent prior to the start of the experiment. Given the lack of any accurate and comparable effect size, I decided to take an approach of sequential analysis to ensuring the study was sufficiently powered while avoiding false positives (Type 1 errors). Sequential analysis involves collecting and analysing data at increasingly large sample sizes while controlling for the Type 1 error rate by effectively lowering the threshold for significance (for an in-depth discussion of the approach, see Lakens, 2014). I planned to collect data in batches of 20 subjects, to analyse that data, and to then decide whether results were convincing enough to conclude that an effect was present and to stop data collection. If no conclusive results were found, I would then continue data collection, but I would halve my threshold p-value for the next iteration (making my threshold for significance in the second batch 0.025, and in the third 0.0125, and so on). From my first sample of 20 participants, three were excluded as their data was incomplete due to issues with recording their data, and I found convincing and statistically significant results from the data of the remaining 17 participants, terminating my sequential analysis plan at the first step. I therefore had a final sample size of 17 participants for this experiment. Participants received £5 per hour as a base rate and could earn an additional bonus payment (up to £3) based on performance in the task.

3.2.1.2 Task Design

Participants performed a skilled motor task where the objective was to throw a ball and hit a target in a 2D aiming challenge. Each trial started with a choice between two balls, and having chosen between the two balls, participants would then throw their chosen ball at a target to complete the trial (see Figure 3-1 – Schematic of procedure followed by individual trials (A) and entire blocks (B). Individual trials consisted of an initial choice between two balls, with full knowledge of the rewards earned under by each option, the size of the target the participant must hit, and the length of the block. Having decided between the two balls based on this information, participants would throw the ball and aim to hit the target. Each block consisted of 5 (short) or 15 (long) trials, after which all rewards earned over all trials would be delivered at once.A). One ball would be orange, the other blue, and participants were instructed in a tutorial that these colours corresponded to two different reward schemes. One ball (which I refer to as the *low-level ball*) would earn more reward (2 points) for each individual hit, while the other (which I refer to as the *high-level ball*) would earn less reward for a hit (1 point) but would, if selected and successfully thrown at the target on *every* trial in a block, triple all the reward earned for that block. Note that the multiplier applied to all reward in a block was a hierarchically

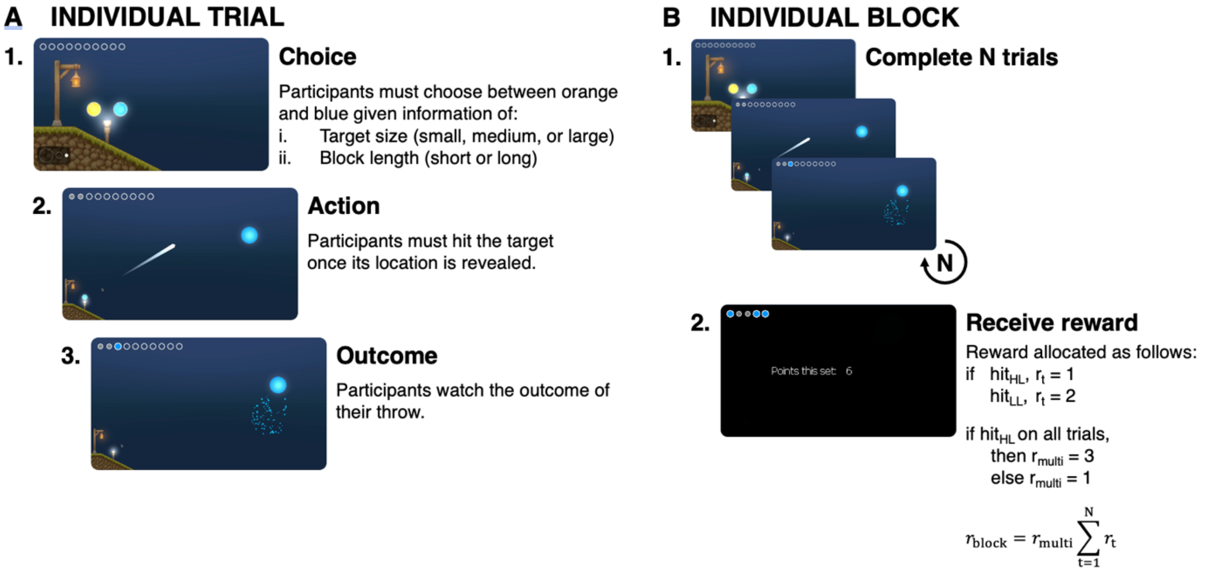

A INDIVIDUAL TRIAL
1.
Choice
Participants must choose between orange and blue given information of:
i. Target size (small, medium, or large)
ii. Block length (short or long)
2.
Action
Participants must hit the target once its location is revealed.
3.
Outcome
Participants watch the outcome of their throw.
B INDIVIDUAL BLOCK
1.
Complete N trials
N
2.
Points this set: 6
Receive reward
Reward allocated as follows:
if hit$_{HL}$, r$_t$ = 1
hit$_{LL}$, r$_t$ = 2
if hit$_{HL}$ on all trials,
then r$_{multi}$ = 3
else r$_{multi}$ = 1
$r_{\text{block}} = r_{\text{multi}} \sum_{t=1}^{N} r_t$

Figure 3-1 – Schematic of procedure followed by individual trials (A) and entire blocks (B). Individual trials consisted of an initial choice between two balls, with full knowledge of the rewards earned under by each option, the size of the target the participant must hit, and the length of the block. Having decided between the two balls based on this information, participants would throw the ball and aim to hit the target. Each block consisted of 5 (short) or 15 (long) trials, after which all rewards earned over all trials would be delivered at once.

higher-level reward than the rewards received for individual hits because the multiplier was contingent upon an aggregate of behaviour over an entire block, whereas individual rewards depended on only one throw. Note also that all rewards were always delivered at the end of a block, irrespective of whether they were individually allocated (low-level balls) or hierarchically allocated (high-level balls) to remove any effect of temporal discounting.

There are two candidate policies of action available in the task: (1) the *low-level* policy ignores the high-level multiplier and maximises low-level reward by selecting the low-level ball on every trial in a block; (2) the *high-level* policy pursues the high-level multiplier and maximises reward by selecting the high-level ball on every trial in a block. Importantly, the high-level policy earns more total reward, but maximum reward here is contingent upon the participant successfully landing *every* shot in a block: any one miss is catastrophic for the maximal pay-out. By contrast, the low-level policy earns less total reward, but a miss on one trial does not detract from the reward earned by hits on other trials. Therefore, to choose between the high- and low-level policies, participants needed to estimate how likely they were to hit the target on each trial in a block. That is, they needed to make a self-efficacy judgment about their ability to hit the target and use this to guide choice. Specifically, if they believed they could consistently hit the target (high self-efficacy) they should have chosen the high-level policy, otherwise they should have chosen the low-level policy.

I manipulated two variables block-by-block to investigate which features of the policies under consideration would influence self-efficacy. First, I varied the difficulty of individual trials by varying the size of the targets participants were aiming to hit. Difficulty was fixed for every trial within a block but could vary between blocks. There were three levels of difficulty: the target could be small, medium, or large, which translated into hard, moderate, or easy difficulties (respectively). Note that, according to Fitts’ Law (Fitts, 1954, 1964), the distance of the target from the ball’s starting position also influences the difficulty of hitting it. While the position of the target was randomly set trial by trial, participants could not see where the targets were while deciding between the two balls, and so their distance from the starting position could not influence choice. However, participants *were* given a cue as to the size of the target at the start of each trial, and so target size could influence choice, with smaller targets being harder to hit. I therefore reasoned that smaller targets would make greater demands on self-efficacy, and would therefore more strongly express self-efficacy effects. Second, I varied the length of each block. There were two levels here: short (5 trials), or long (15 trials). This manipulation made the high-level policy more difficult to execute as the likelihood of hitting *all* targets in a block decreases exponentially with the number of trials in a block. The 2 x 3 factorial design I used here therefore allowed us to investigate how classical notions of self-efficacy interact with the features of high-level behaviour. Specifically, I predicted that as either block length or trial difficulty increased, perceived ability to carry out the high-level policy would decrease, under the hypothesis that people are sensitive to the length of the high-level policies they choose to perform when estimating their ability to carry them out.

Given the high-level nature of the candidate behaviours in the task, the most informative decisions for my hypotheses were made on the first trial of each block. At the start of every trial, subjects are given two cues: (1) one cue for the difficulty of all trials in that block; and (2) another cue for the number of trials in that block. These cues of difficulty and length gave my participants all the information they needed to make an informed self-efficacy judgment for an entire block on the first trial and to then use that judgment to decide whether or not to attempt to perform the more rewarding but more difficult high-level policy. A decision to attempt this would be indicated by selecting the high-level ball on the first trial, and such a decision would be equivalent to answering “yes” to the question “are you likely to succeed in performing this behaviour?” (or, more colloquially, “can you cope with this?”) as per Bandura's (1977) original guidelines for measuring self-efficacy. I planned to direct most of my analyses at these first trials in each block, with my central hypothesis being that I would observe less frequent commitment to the high-level policy on these trials as difficulty and block length increased. That is, participants would correctly perceive that sustained effort was likely to tax their performance capacity.

3.2.1.3 Procedure

Having signed up for the experiment via Prolific (www.prolific.co), all participants would begin by providing informed consent. They would then complete a brief tutorial which introduced them to the skilled motor task (with a practice throw for each target size), and the reward schemes were then explained in full. Participants were told exactly how many points could be earned by selecting either of the two balls and were then asked to complete two in-progress blocks until they had earned as much as possible from each block. The first of these in-progress blocks had six of ten trials already completed by selecting the low-level ball, and the second had six of ten trials already completed by selecting the high-level ball. To earn maximal pay-outs for each block, participants would need to select the low-/high-level ball and hit the target on the remaining four trials. This provided the participants with direct experience of the maximum reward available under each of the low- and high-level policies. To end the tutorial, participants answered four questions to gauge their understanding. They were asked (1) how many different target sizes they could encounter, (2) how many points they would earn for hitting the target with the low-level ball, (3) how many points they would earn for hitting the target with the high-level ball, and (4) which of the low- and high-level policies would earn more reward given perfect performance over a block. Participation in the full experiment was permitted *only* if a correct answer was given for all four questions to ensure that participants understood the contingencies involved. Having completed the tutorial, participants then moved into the full experiment. My 2 x 3 factorial design provided six combinations of trial difficulty and block length, which were repeated five times for a total of 30 blocks. Each participant would complete the 30 blocks in a random order. Each block began with an inter-block screen which waited for input from the participant, which when provided would initiate the trials. Trials began with a choice between balls, which was followed by the skilled task of throwing the ball at the target, and trials ended with observation of the result (see Figure 3-1 – Schematic of procedure followed by individual trials (A) and entire blocks (B). Individual trials consisted of an initial choice between two balls, with full knowledge of the rewards earned under by each option, the size of the target the participant must hit, and the length of the block. Having decided between the two balls based on this information, participants would throw the ball and aim to hit the target. Each block consisted of 5 (short) or 15 (long) trials, after which all rewards earned over all trials would be delivered at once.A). Each trial transitioned directly into the next. Once each trial in a block had been completed, all rewards were then delivered at once (see Figure 3-1 – Schematic of procedure followed by individual trials (A) and entire blocks (B).

Individual trials consisted of an initial choice between two balls, with full knowledge of the rewards earned under by each option, the size of the target the participant must hit, and the length of the block. Having decided between the two balls based on this information, participants would throw the ball and aim to hit the target. Each block consisted of 5 (short) or 15 (long) trials, after which all rewards earned over all trials would be delivered at once.B). Having completed all blocks, participants were debriefed and then redirected to Prolific to terminate the experiment.

3.2.1.4 Analysis

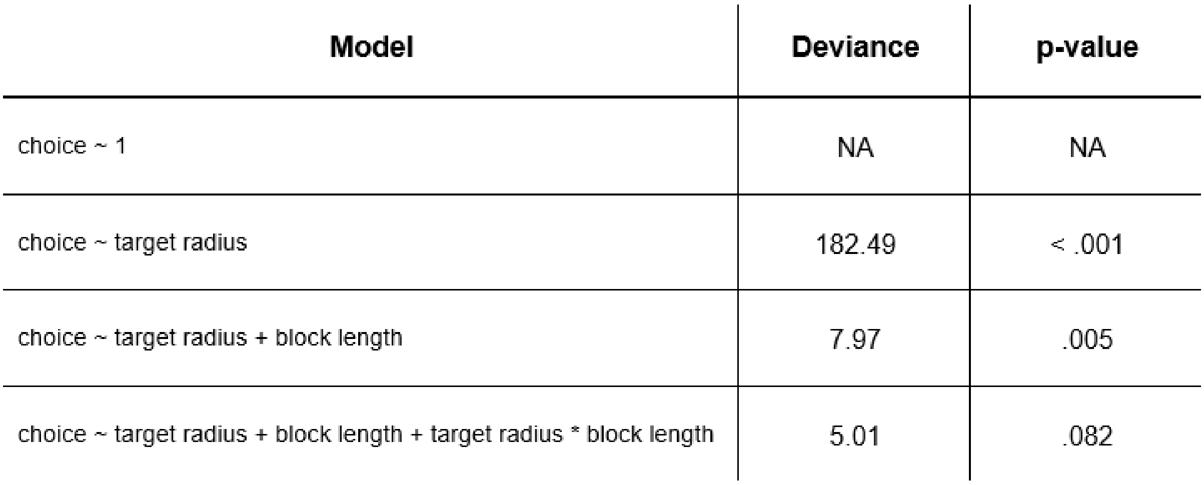
As mentioned above, I planned to focus my analyses on the first trial of each block, as it was here where commitments to one of the two policies would be made. I had two variables of interest here: (1) choices; and (2) response times. For choice data, I fit nested logistic regression models to choice data to evaluate whether including block length, trial difficulty, and the interaction between these two predictors led to significant improvements in model fit. For response time data, I had two epochs of interest. First, I measured the time taken to decide between the two balls, which I named decision time (DT). Second, I measured the time interval between the decision of which ball to throw, and the time of actually throwing it. I called this action time (AT). I planned to analyse DT by investigating how DT changed with different trial types. I had four trials of interest: (1) *initial* trials were the first trials in each block, which, as discussed, are the trials where commitments were made to one of the two policies of action; (2) *high-level* trials were trials in a block that carried out a commitment to the high-level policy; (3) *first miss* trials were the first trials in a block where a participants had decided to perform the high-level policy but missed the target; and (3) *low level* trials were trials in a block that carried out a commitment to the low-level policy. I planned to compute the average DT for each participant in each of these trial types and again to analyse these using the Kruskal-Wallis test (given non-normality). For AT, I planned to compute the average AT for each combination of trial difficulty and block length separately for trials in which subjects chose the high-level and low-level balls, and to analyse these data in a 2x2x3 repeated measures ANOVA on log AT data (log AT satisfied the normality requirement of ANOVA; results from Shapiro Wilk test were W = 0.99, p = 0.317).

3.2.2 Results

3.2.2.1 Choices

As measured by choices on the first trial of every block, willingness to commit to the difficult high-level policy of action towards maximising reward diminished with increased trial difficulty (see row 2 of Table 3-1 – results of analysis of deviance tests to evaluate incremental improvements in model fit of a sequence of nested logistic regression models. The nested models successively introduce target radius, block length, and the interaction between these two predictors as predictors of choice behaviour. Each row presents the results of a deviance test for the model of that row against the model in the row above it.) and increased block length (see row 3 of Table 3-1 – results of analysis of deviance tests to evaluate incremental improvements in model fit of a sequence of nested logistic regression models. The nested models successively introduce target radius, block length, and the interaction between these two predictors as predictors of choice behaviour. Each row presents the results of a deviance test for the model of that row against the model in the row above it.). Rather unsurprisingly, participants were sensitive to increases in individual trial difficulty and to the exponential increases in difficulty that came with repeating trials multiple times (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.A). I also found that including the interaction between trial difficulty and block length as a predictor of choice led to a marginal improvement in model fit (see row 4 of Table 3-1 – results of analysis of deviance tests to evaluate incremental improvements in model fit of a sequence of nested logistic regression models. The nested models successively introduce target radius, block length, and the interaction between these two predictors as predictors of choice behaviour. Each row presents the results of a deviance test for the model of that row against the model in the row above it.). Inspection of the proportions of high-level commitments made by participants (see Figure 3-2) indicates that this was driven by a highly significant boost in willingness to commit to the high-level policy for short rather than long block lengths for trials of moderate difficulty (t(16) = 4.60, *p* < .001, d = 1.12), but not for trials of easy (t(16) = 1.46, *p* = .164) nor hard (t(16) = 1.10, *p* < .290) difficulties. This non-linearity suggests that participants were sensitive

Table 3-1 – results of analysis of deviance tests to evaluate incremental improvements in model fit of a sequence of nested logistic regression models. The nested models successively introduce target radius, block length, and the interaction between these two predictors as predictors of choice behaviour. Each row presents the results of a deviance test for the model of that row against the model in the row above it.



Model	Deviance	p-value
choice ~ 1	NA	NA
choice ~ target radius	182.49	< .001
choice ~ target radius + block length	7.97	.005
choice ~ target radius + block length + target radius * block length	5.01	.082

to the non-linear and exponential decreases in the likelihood of hitting all shots in a block that occurred as the length of the block increased. In short, the effect of moderate difficulty is much more pronounced in a lengthier block.

The changes I observed in willingness to commit to the high-level policy tracked changes observed in performance accuracy (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block

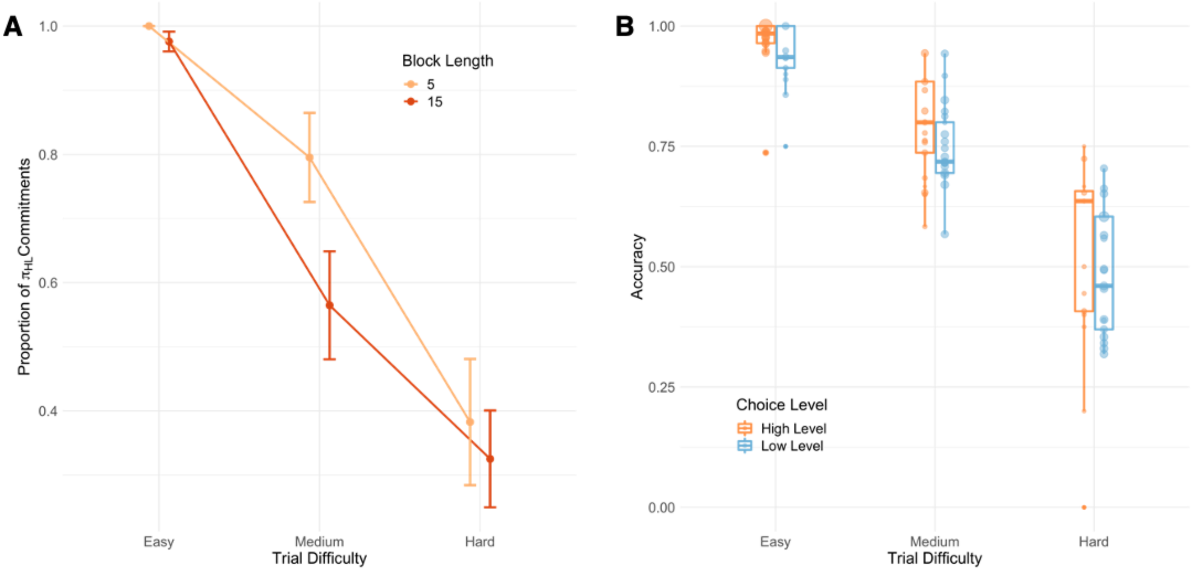


Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.

length. (B) accuracy of all trials under each choice level split by trial difficulty.B). Here, I applied the same analysis approach of tested a nested set of logistic regression models but now to predict performance accuracy. I unsurprisingly found that including trial difficulty as a predictor of performance accuracy led to a highly significant improvement in model fit (deviance = 70.01, p < .001), which is reflected in the observed decrease in willingness to commit to the high-level policy with increased trial difficulty. Second, although it is difficult to analyse these data given that participants selected the high-level ball only infrequently for harder difficulties (making accuracy difficulty to reliably compute), an inspection of accuracy data split by choice level suggests that low-level choice trials were slightly less accurate than high-level choice trials (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.B). Although this is speculative, this would be a sensible strategy, as the contingencies involved in the high-level policy place greater weight on each individual hit, perhaps encouraging more cognitive effort and a slight increase in accuracy.

3.2.2.2 Response Times

I analysed two distinct sets of response times: (1) log decision time (DT) was the log of the time taken to decide between the two balls on each trial; and (2) log action time (AT) was the log of the time taken to then throw the chosen ball at the target. I analysed DT separately for four distinct trial types to test my assumption that the first trial in a block was the time at which a high-level commitment was made to one of the two policies of action under consideration. Indeed, I found a significant effect of

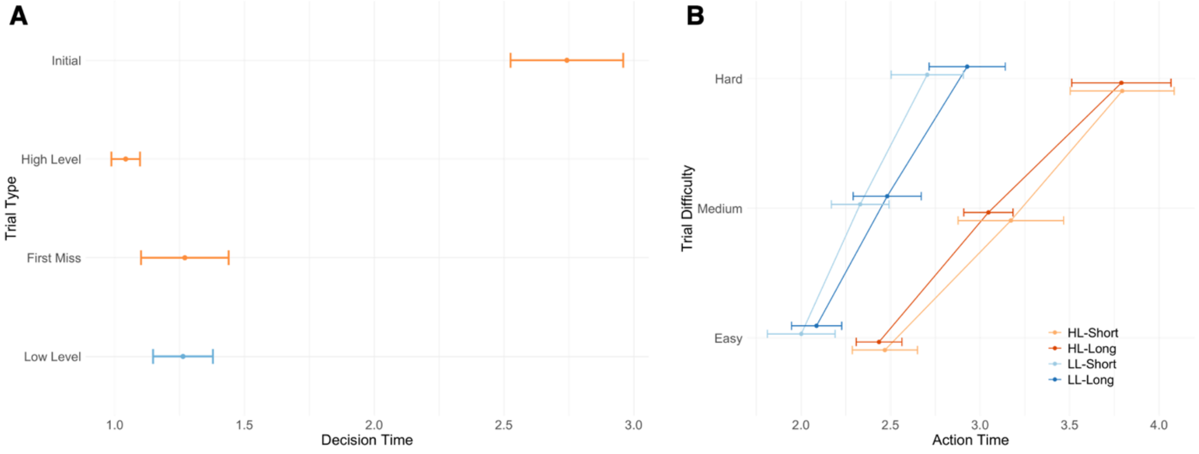


Figure 3-3 – Response times split by (A) decision time, which is the time taken from the beginning of each trial to decide between high- and low-level balls; and (B) action time, which is the time taken from the moment of choosing a ball to throw it.

trial type on decision time ($\chi^2(3)$ = 32.27, p < .001), with decision times on the initial trial in a block being substantially slower than any other trial type (see Figure 3-3 – Response times split by (A) decision time, which is the time taken from the beginning of each trial to decide between high- and low-level balls; and (B) action time, which is the time taken from the moment of choosing a ball to throw itA). This lines up with the hierarchical description I provide for behaviour. A hierarchical account claims that decisions between policies are made on the first trial, and that all following decisions follow deterministically from the initial commitment to a policy. Deliberations between the two policies under consideration in light of trial difficulty and block length would take time and would under this hierarchical account occur only on the first trial in a block. Therefore, this deliberative process can account for the marked increase in decision time on the first trial in a block, and much shorter decision times in all trials that follow.

I analysed action times by considering how they changed with trial difficulty, block length, and choice level. I hypothesised that the time taken to make an action would increase with the difficulty of succeeding, and that the choice between high- and low-level balls would also influence action time as there is more pressure to succeed on each individual shot under the high-level policy, which might encourage participants to take their time. I found a significant main effect of trial difficulty (F(2, 32) = 19.65, *p* < .001, η^2_{G} = 0.07), with action times increasing as trials became more difficult (see Figure 3-3 – Response times split by (A) decision time, which is the time taken from the beginning of each trial to decide between high- and low-level balls; and (B) action time, which is the time taken from the moment of choosing a ball to throw itA). Unfortunately, given that participants were free to decide whether to perform the high-level policy on each block, and given that most participants opted not to perform the high-level policy for lengthy and difficult blocks (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.), I do not have enough data for all block types to include choice level in my analysis. Nevertheless, an inspection of the sparse data I do have reveals that action times were substantially slower under the high-level policy than under the low-level policy (see Figure 3-3 – Response times split by (A) decision time, which is the time taken from the beginning of each trial to decide between high- and low-level balls; and (B) action time, which is the time taken from the moment of choosing a ball to throw itB). This is speculative, but it does lend further support to the notion that participants took greater care and applied more effort to succeeding under the high-level policy, which lines up with the catastrophic consequences of failing.

3.2.3 Discussion

In this first experiment, I found that willingness to commit to a high-level policy of action was dependent upon both the length of the policy and the difficulty of the individual actions it prescribed. Participants were less likely to attempt to perform the high-level policy in lengthier blocks and in blocks with more difficult trials, and their choices were sensitive to the exponential decreases in the probability of success that come with lengthier blocks of more difficult trials. My hierarchical interpretation of the behaviours on display here was supported by the result that participants took far longer to decide what to do on the first trial in a block than on any other trial, indicating that they were at that time deciding on a course of action that they would follow for the rest of the block (Kaller, Unterrainer, & Stahl, 2012). There were also indications that participants exerted more effort on trials where they engaged with the high-level policy; accuracy was slightly (though not significantly) higher on these trials, and the time taken to execute action was lengthier on these trials than on trials where participants pursued low-level reward. Although these final results are speculative, they do match the contingencies involved: a single failure under the high-level policy diminishes the reward earned by all previous successes, whereas failing on any one trial under the low-level policy does not affect the rewards earned by hits on other trials. In sum, self-efficacy theory (Bandura, 1974, 1977; Kirsch, 1995) can and should be extended to consider the features of hierarchically organised behaviour.

I found an impressive correspondence between the willingness of the participants to engage with the high-level policy and the decreases in likelihood of success as the length of that policy increased, indicating that people in general make accurate hierarchical self-efficacy judgments. These decreases in the likelihood of success with longer sequences of actions are exponential. This non-linear decay in the probability of success was reflected in a significant interaction between the difficulty of the individual actions and the number of times they were to be repeated. For short block lengths, moving from easy to moderately difficult trials resulted in only a modest drop in willingness to commit to the high-level policy. By contrast, for long block lengths, moving from easy to moderately difficult trials resulted in a severe decrease in this same measure. Therefore, the same shift in difficulty had a much more pronounced effect on choice for long block lengths than for short ones, and this can be attributed to the exponential decreases in the likelihood of success that come with increases in length. My participants were therefore not only vaguely aware that lengthier policies of action would be more difficult to perform, but they were in fact precisely aware of the non-linear decreases in the prospect of success that came with these lengthier sequences. Therefore, self-efficacy judgments were made not based only on the perceived difficulty of any individual behaviour as described by classical self-efficacy theory (Bandura, 1977), but also very precisely on the impact of needing to repeat difficult actions many times.

In experiment 3-1, I successfully extended self-efficacy to include length as a feature of high-level action, and my novel approach of measuring self-efficacy judgments using choice as a proxy proved successful. However, the results here do not speak to the second of my two hypotheses, namely that length influences choice not only by making the prospect of repeated success less likely but also by the aversive prospect of restricting choice. I propose that a commitment to a lengthy course of action might be aversive because that lengthy course of action would limit and restrict future decisions, and in a dynamic and ever-changing world this might mean turning down options of which people are currently unaware. In my first experiment, the environment was static: rewards were fixed and fully explained at the beginning of the experiment. Whilst this was useful for a clean test of an extension of self-efficacy theory to hierarchically organised behaviour, there was no need for my participants to fear any future restrictions on choice (due to the static nature of the rewards in the task). To test the hypothesis that restrictions on future choice are aversive, I modified my design in a second experiment to include dynamic changes in reward that could provide an incentive to avoid restrictions on choice. I structured rewards such that for easy trials, the high-level policy remained the most rewarding policy over a long-run average, and so I asked whether my participants would deviate from the optimal high-level policy on these trials towards avoiding tying the hands of their future selves.

3.3 Experiment 3-2

3.3.1 Methods

3.3.1.1 Participants

This second experiment was approved by the UCL Research Ethics Committee, and it was also hosted online. Participants were again recruited via Prolific (www.prolific.co), and all were then redirected to a personal website where the experiment was hosted. All participants provided informed consent prior to the start of the experiment. I followed the same sequential analysis approach as described for the previous experiment here, however in anticipation of a weaker effect size I increased the size of each intermittent sample to 30 (for experiment 3-1, this was 20). Otherwise, I again planned to collect data in batches, to analyse the data, and to then decide whether results were sufficiently convincing to terminate data collection. If no conclusive results were found, I would collect an additional sample of 30 participants, but I would halve my threshold p-value for each successive batch. To avoid issues of attrition, I collected an initial sample of 35, of whom 33 completed the entire experiment. From this first sample of 33 participants, I found convincing statistically significant results and so I terminated data collection at this first step. Participants received £5 per hour as a base rate and could earn an additional bonus payment (up to £3) based on performance.

3.3.1.2 Design & Procedure

Experiment 3-2 followed the same design and procedure as experiment 3-1 but for two changes. First, and most importantly, I introduced a third, white ball, which could, replace the low-level ball and offer a substantially larger low-level reward if chosen. This would happen only infrequently and was unpredictable. I refer to this as the *boost*. The boost could not appear on the first trial in a block, to maintain the need for participants to decide between low- and high-level balls and their associated policies, but it would otherwise appear stochastically with probability 1/14 on every other trial.

This boost introduced variability to the environment, and importantly it provided an incentive to not want to restrict choice for an entire block by committing to the high-level policy. If a boost might happen, an extra reward might be available, but pursuing it was possible only if one had not already committed to the high-level policy, or if one was willing to abandon a commitment to the high-level policy. Deviating from the high-level policy to take the boost would result in a loss of the bonus available under that policy, and so participants needed to trade off the value of the bonus against the possibility of the boost. I set the magnitude of the boost to a value of 10, which kept the high-level policy as the optimal behaviour in the long run for easy blocks (see the next section for computations of expected value).

The second change made was to the number of repetitions of each combination of block length and trial difficulty. I again included short and long block lengths and easy, moderate, and hard difficulties. However, I reasoned that of these conditions, I would be most likely to observe the effect of aversion to choice restriction on behaviour on short blocks with easy trials. For moderate and hard difficulties and for long blocks, the difficulty of executing the high-level policy successfully was likely to be the dominant dissuading factor influencing commitment to it. Therefore, to focus my measurements on any effect of aversion to choice restriction, I doubled the number of easy and short blocks, as this is where I expected to observe this effect most clearly. I also lowered the number of total repetitions of all factor combinations to two. These two changes to the design of experiment 3-1 allowed us to investigate whether participants avoided engaging with optimal high-level policies, specifically because doing so involved an aversive restriction on future choices.

3.3.1.3 Analysis

I planned to analyse the data from experiment 3-2 initially in much the same way as I analysed behaviour in experiment 3-1. I analysed the proportion of commitments to the high-level policy over all initial trials for all blocks using nested logistic regression models to investigate whether I could replicate my findings from experiment 3-1. I also tested whether willingness to commit to the high-level policy on blocks with *easy* trials differed between the two experiments, again using nested logistic regression models. This would test whether introducing the boost (in experiment 3-2) resulted in a lesser willingness to commit to the high-level policy (relative to experiment 3-1). To further investigate whether my participants did indeed deviate from the optimal solution to the task, I computed the normative expected value (EV) of each policy given a full range of accuracy levels. For each of the EV computations for low- and high-level policies, see the equations below. Parameters in these equations are as follows: $\mathrm{EV}_{\pi_{\mathrm{x}}}$ describes the expected value under policy π_x; $\mathrm{Pr}(hit)$ describes the probability of hitting the target on a given trial; $\mathrm{Pr}(boost)$ describes the probability of encountering a boost on a given trial; r_{boost} describes the reward offered by the boost; r_{LL} describes the reward offered by the low-level ball; r_{multi} describes the multiplier applied to all reward earned if the participant chooses and successfully hits all high-level targets in a block; and r_{HL} describes the reward offered by the high-level ball on a single trial.

$$\mathrm{EV}_{\pi_{\mathrm{LL}}} = \sum_{t=1}^{\mathrm{T}} \mathrm{Pr(hit)} \cdot \left(\mathrm{Pr(boost)} \cdot r_{\mathrm{boost}} + \left(1 - \mathrm{Pr(boost)}\right) \cdot r_{\mathrm{LL}}\right)$$

$$\mathrm{EV}_{\pi_{\mathrm{HL}}} = \mathrm{Pr(hit)}^{\mathrm{T}} \cdot r_{\mathrm{multi}} \sum_{t=1}^{\mathrm{T}} \mathrm{Pr(hit)} \cdot r_{\mathrm{HL}} + (1 - \mathrm{Pr(hit)}^{\mathrm{T}}) \cdot r_{\mathrm{multi}} \sum_{t=1}^{\mathrm{T}} \mathrm{Pr(hit)} \cdot r_{\mathrm{HL}}$$

These expected value computations provide a normative estimate of the relative values for a given accuracy level ($\mathrm{Pr(hit)}$) and block length (T). For each participant, I could record three distinct accuracy levels (one each for the three trial difficulties). I took these accuracy levels and computed the expected value of each policy given these accuracies separately for each of the two block lengths. I could then take the net EV in favour of the high-level policy, and this provided us with a normative guide for choice: if net EV was positive, the optimal choice for maximising reward in the long run was the high-level policy; if net EV was negative, the optimal choice for maximising reward was the low-level policy. I planned to compute these net EV values for each participant in both experiments and to test whether the correlation between net EV and willingness to commit to the high-level policy was diminished in experiment 3-2 relative to experiment 3-1 (by testing the linear model given by $\mathrm{prop_{HL}} \sim \mathrm{net\ EV} + \mathrm{boostAvailable} + \mathrm{net\ EV} * \mathrm{boostAvailable}$). This would test whether introducing variability in the rewards available prompted deviations from optimal behaviour, thereby making behaviour less dependent on EV. In essence, this tests whether the prospect of a boost in reward for low-level choices dissuaded commitment to a high-level policy that could not pursue that boost. Such a result would strongly suggest that high-level policies were not chosen, even when they were attractive, because they effectively tie the hands of the agent and foreclose future possibly tempting action choices.

3.3.2 Results

3.3.2.1 Choices

I found that the willingness of my participants to commit to the high-level policy of action in experiment 3-2 was predicted by trial difficulty and block length in much the same way as observed in experiment 3-1 (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.**Error! Reference source not found.**A). I found that including target radius (see row 2 of Table 3-2) and block length (see row 3 of Table 3-2) led to significant improvements in logistic regression model fit to choice behaviour. Including the interaction between these two predictors however did not significantly improve model fit (see row 4 of

Table 3-2 – results of analysis of deviance tests to evaluate incremental improvements in model fit of a sequence of nested logistic regression models. The nested models successively introduce target radius, block length, and the interaction between these two predictors as predictors of choice behaviour. Each row presents the results of a deviance test for the model of that row against the model in the row above it.

Model	Deviance	p-value
choice ~ 1	NA	NA
choice ~ target radius	117.59	< .001
choice ~ target radius + block length	12.13	< .001
choice ~ target radius + block length + target radius * block length	2.06	.357

Table 3-2), indicating that participants in this experiment were not as sensitive to the exponential decreases in the probability of success that come with increased block length. I successfully replicated the key result of experiment 3-2, which was a sensitivity to the length of an action sequence while making judgments about the likelihood that one could successfully execute that sequence. I find modest evidence

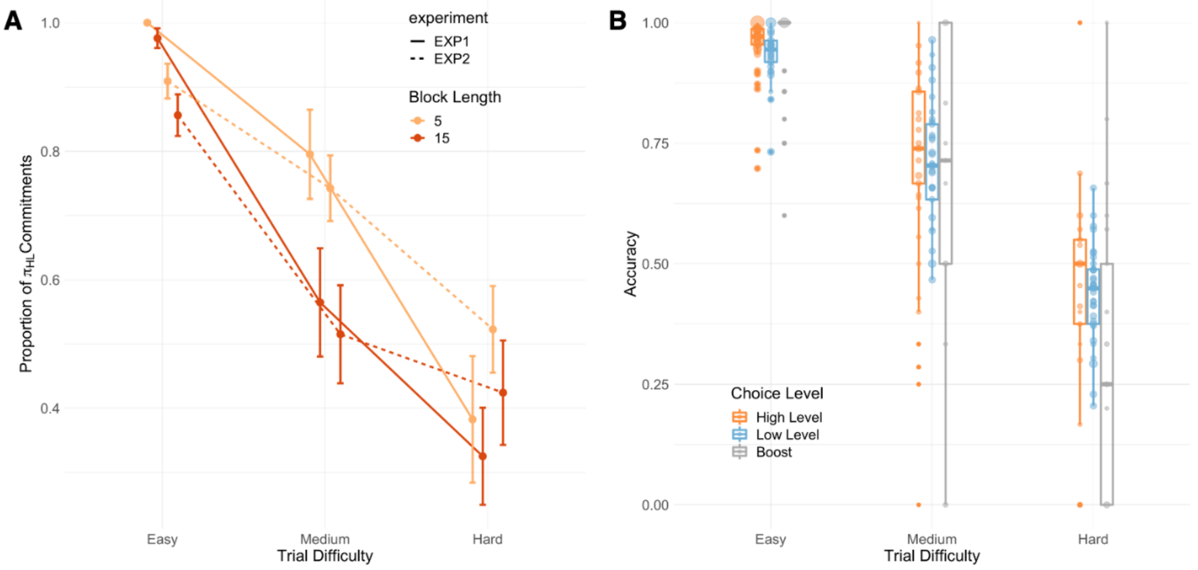


Figure 3-4 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty, block length, and experiment; (B) accuracy of all trials under each choice level split by trial difficulty.

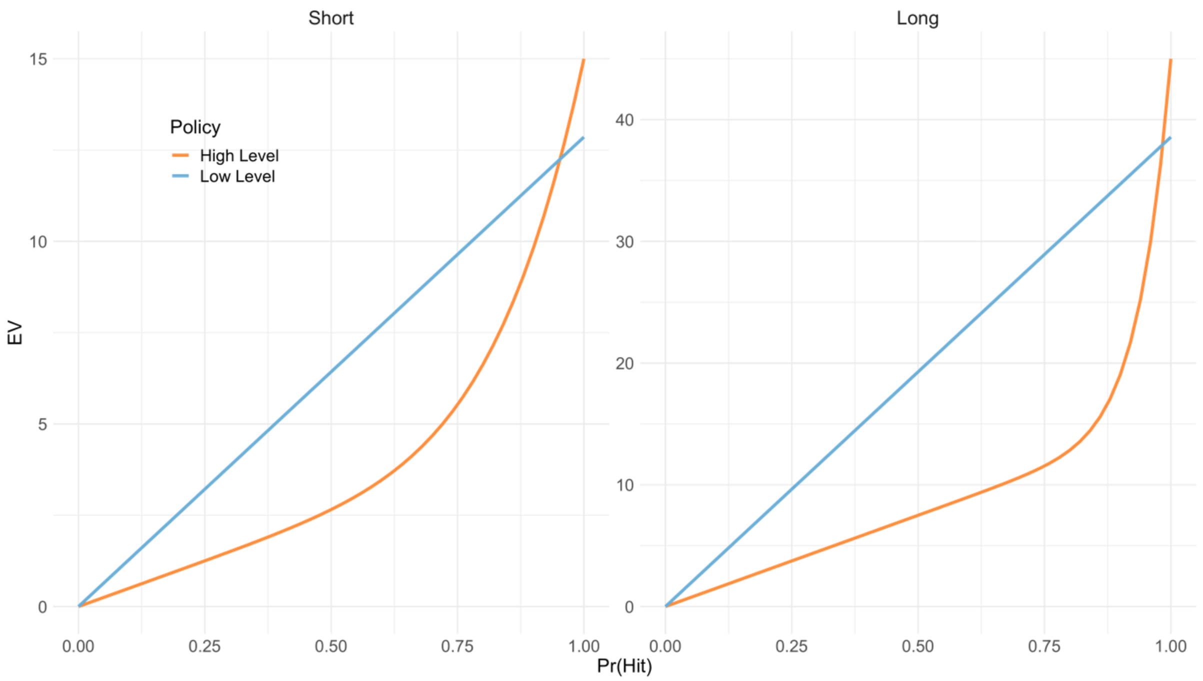


Figure 3-5 – Expected value for each policy given the two different block lengths and the full range of accuracy levels (Pr(Hit)).

in favour of a sensitivity to changes in this likelihood with the length of a behaviour, as was observed in the interaction between trial difficulty and block length in experiment 3-1.

Next, I tested whether commitments changed between experiments and over different block lengths for easy trials alone. To do this, I added block length and experiment as predictors to a logistic regression model that aimed to predict choice behaviour on easy trials. I found no significant improvement in model fit when block length was introduced (deviance = 1.43, p = 0.231), but a highly significant improvement in fit where experiment was introduced (deviance = 21.78, p < .001). If we inspect the accuracy of participants on these trials over both experiments, most were at ceiling (see Figure 3-2 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty and block length. (B) accuracy of all trials under each choice level split by trial difficulty.B & Figure 3-4 – (A) proportion of all blocks where participants chose the high-level ball on the first trial in the block, indicating a commitment to the high-level policy, split by trial difficulty, block length, and experiment; (B) accuracy of all trials under each choice level split by trial difficulty.B). Therefore, although the participants in experiment 3-2 would in all likelihood have successfully executed the sequence of actions required under the high-level policy given their accuracy levels, they were significantly less likely to commit to the high-level policy in the presence of a possible boost in low-level value than were the participants in experiment 3-1 where no such boost was available. In sum, where a tempting prospect of future low-level reward was present, participants were less likely to engage with a high-level policy of action.

3.3.2.2 Commitment and Expected Value

To analyse whether the participants' behaviour deviated from the optimal solution to the task when faced with the prospect of a low-level boost in value, I computed the expected value (EV) of each policy for the full range of accuracy levels and for each of the two block lengths included in my experiments (see 3.3.1.3 Analysis). The EV curves for different accuracy levels can be found in Figure 3-5. At high, near perfect accuracy, the optimal choice for maximising long-run reward is to opt for the high-level policy, but for lower accuracies, the optimal choice is to opt instead for the more forgiving low-level policy. Note that the exponential decreases in the probability of successfully hitting all shots in progressively lengthier sequences is reflected here in the EV of the high-level policy.

I mapped the accuracy level of each participant for each level of trial difficulty and block length onto these EV curves and computed six net EV scores ($\mathrm{EV}_{\pi_{\mathrm{HL}}} - \mathrm{EV}_{\pi_{\mathrm{LL}}}$) per participant (one each for every combination of trial difficulty and block length). These net EV scores provide a normative guide for choice: where net EV is

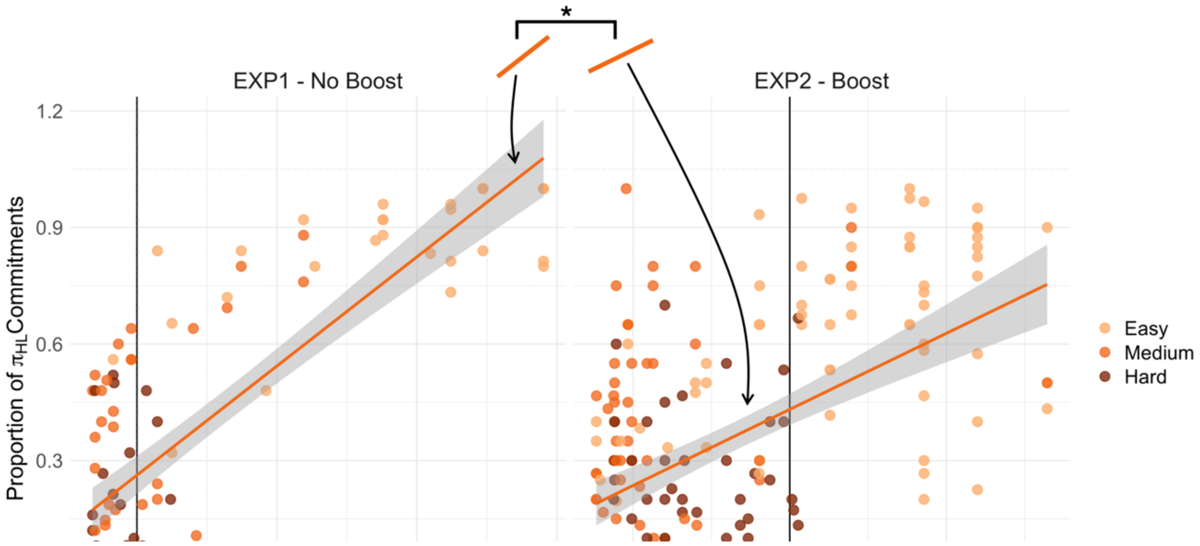
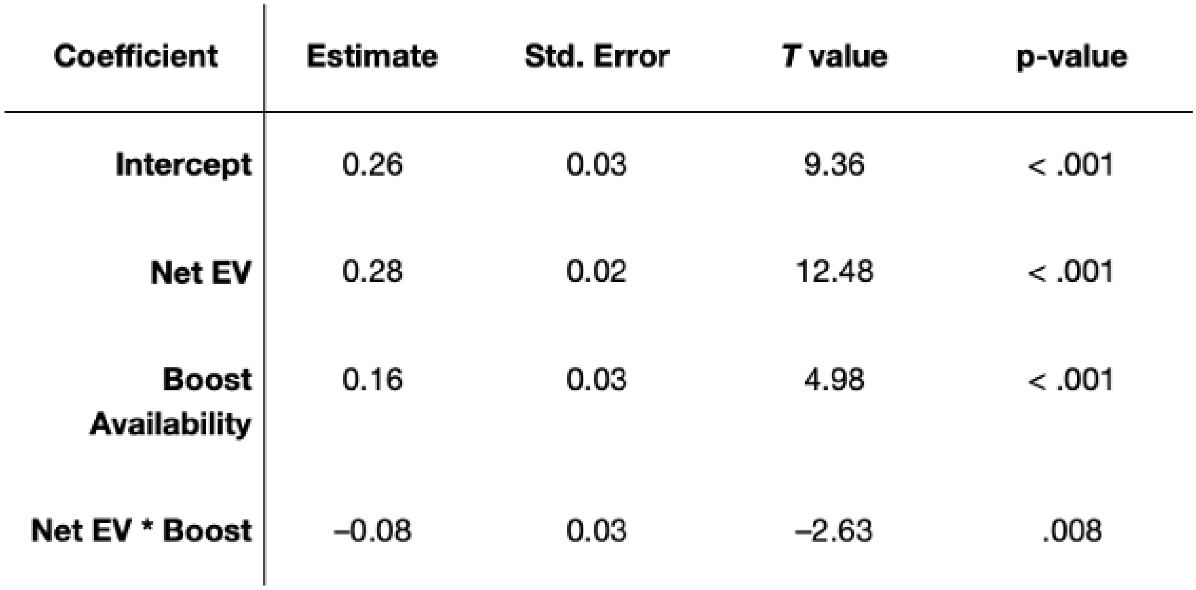


Figure 3-6 – Correlation between net expected value (Net EV) in favour of the high-level policy (computed as EV for high-level policy – EV for low-level policy) and the proportion of high-level commitments made by each participant for each difficulty level (which is associated with a different accuracy level and therefore different net EVs) separately for each experiment. The difference in slopes between the two experiments is significant, with the presence of a low-level boost in resulting in a significantly shallower slope, indicating that decisions here deviated further from the optimal solution.

Table 3-3 – Regression coefficients for the linear model given by: $prop_{HL} \sim net\ EV + boost\ availability + net\ EV * boost\ availability$



Coefficient	Estimate	Std. Error	*T* value	p-value
Intercept	0.26	0.03	9.36	< .001
Net EV	0.28	0.02	12.48	< .001
Boost Availability	0.16	0.03	4.98	< .001
Net EV * Boost	–0.08	0.03	–2.63	.008

positive, the optimal choice is to select the high-level policy; and where net EV is negative, the optimal choice is to select the low-level policy. I used a linear model to predict the proportion of high-level commitments (as measured by choice on the first trials of each block) from these net EV scores and from whether the boost was available (i.e., whether the participants completed experiment 3-1 or experiment 3-2). The interaction term between net EV and boost availability would tell us whether there was any difference in the relationship between behaviour and the optimal estimate of the value of the two policies between the two experiments. The model predicted a significant portion of the variance in high-level commitments ($\mathrm{F}(6, 293) = 109.00, p < .001,\ \mathrm{R}^2 = 69.10\%$). Further, not only did I find that net EV and the availability of a boost predicted commitment alone, but I also found that the interaction between these two predictors accounted for a significant proportion of the variance in willingness to commit to the high-level policy (see Table 3-3 – Regression coefficients for the linear model given by: $prop_{HL} \sim net\ EV + boost\ availability + net\ EV * boost\ availability$). This interaction term translated into a shallower correlation between net EV and the proportion of high-level commitments for experiment 3-2 (*r* = 0.50) than for experiment 3-1 (*r* = 0.81) (see Figure 3-6 – Correlation between net expected value (Net EV) in favour of the high-level policy (computed as EV for high-level policy – EV for low-level policy) and the proportion of high-level commitments made by each participant for each difficulty level (which is associated with a different accuracy level and therefore different net EVs) separately for each experiment. The difference in slopes between the two experiments is significant, with the presence of a low-level boost in resulting in a significantly shallower slope, indicating that decisions here deviated further from the optimal solution.). In sum, where there was a prospect of an infrequent but large boost in low-level reward (as in experiment 3-2), participants would deviate further from the optimal solution to the task. This suboptimality took the specific form of not tying themselves to a course of action that would prevent them from taking advantage of that boost. Where no such prospect was present (as was the case in experiment 3-1), participants followed more closely the optimal solution to the task. Thus, improbable but attractive rewards significantly detracted from participants' willingness to adopt a high-level policy. The possible lucky bonus significantly undermined the long view.

3.3.3 Discussion

In this second experiment, I found that people were less willing to commit to lengthy high-level policies of action in dynamic environments than in static ones. This was the case even when opting for a low-level (rather than high-level) policy in fact lead to less total reward in the long run. I computed the expected value of each option in the task and found that decisions were less dependent on expected values where rewards in the environment were variable. Rational and normative maximisation of value conventionally recommends choosing the action of highest expected value. Therefore, this normative model was insufficient to explain behaviour. It seems my participants were happy to deviate from the optimal solution to the task to avoid tying their hands to a high-level policy of behaviour that would limit their future decisions. In particular, they avoided choices that precluded reaping the advantage of a future lucky occurrence (the boost) that was unlikely but attractive.

It is unclear whether this strategy of avoiding future restrictions on choice is, strictly speaking, a component of self-efficacy. Early conceptions of self-efficacy were concerned primarily with perceived ability to successfully execute a target behaviour (Bandura, 1974, 1977, 1984; Kirsch, 1995), such as successfully throwing a ball and hitting a target - a paradigm I have reprised here. The subsequent development of *coping* self-efficacy enriched the theory to include perceived capability to deal with barriers that arise during the maintenance of a behaviour (Binsch, Wabeke, Koot, Venrooij, & Valk, 2016; Chesney, Neilands, Chambers, Taylor, & Folkman, 2006; DiClemente, Fairhurst, & Piotrowski, 1995; Kirsch, 1995; Schwarzer & Renner, 2000; Williams, 1995). While coping self-efficacy has not been explicitly tied to hierarchically organised action, coping self-efficacy seems like an inherently hierarchical concept; since it captures the belief that one will be able to maintain a lengthy policy of action despite any challenges that arise. The present result places on a computational footing, for the first time, the intuition that self-efficacy involves predictions of one's biases (in this case towards high-level policies) of future choice behaviour.

There is close contact here with failures of self-regulation in addiction (Sayette, 2004). Indeed, my task captures a widely acknowledged property of addictive behaviours, namely that a single deviation from the high-level policy results in a complete failure to achieve the outcome to which the high-level policy is directed. The alcoholic has only to open the bottle in order to fall (Ibsen, 1891). Not only does my task share the same contingencies of addicts' relapse behaviours, but both cases moreover share an essential feature: maintaining the high-level policy involves self-restraint and cognitive effort (which has been demonstrated to influence preferences, Croxson, Walton, O'Reilly, Behrens, & Rushworth, 2009) to resist taking an immediate and highly rewarding off-policy alternative.

Given my results and my hierarchical lens, I suggest a novel and more formal definition for coping self-efficacy. Self-efficacy is the perceived capability to maintain adherence to a high-level policy to which people may be genuinely committed, even in the face of new and more immediately rewarding alternative courses of action. Self-efficacy requires not only that one knows that one can do the right thing in the current choice nexus. It also requires that one believes this choice preference is resilient to tempting but EV-neutral (or EV-damaging) reward possibilities. True self-efficacy thus involves an element of resilience to external manipulations of the reward landscape.

3.4 General Discussion

Across two experiments, I have provided a systematic treatment of how self-efficacy theory can be extended to describe self-efficacy judgments made about hierarchically organised behaviours. I focussed on the length of a high-level behaviour (that being the number of lower-level actions it prescribes) as the most relevant feature of hierarchically organised behaviour for self-efficacy, and I identified two ways in which self-efficacy judgments (as measured by choice) are influenced by the length of a high-level policy of action. First, as the length of a behaviour increases the likelihood of successfully executing all actions it prescribes decreases exponentially, and I found in experiment 3-1 that my participants were sensitive to this exponential decrease. Second, as the length of a behaviour increases it requires that more future decisions adhere to the actions it prescribes, and I found in experiment 3-2 that my participants were averse to this restriction on choice. I have therefore identified two ways in which length as a central feature of hierarchical action influences self-efficacy judgments and as a result influences choice between actions.

Most applications of self-efficacy theory draw on the flat (i.e., non-hierarchical) descriptions of behaviour included in the theory (Bandura, 1977) to design measures that will estimate perceived capability to execute *single* actions in isolated contexts. To refer to an earlier example, the smoking abstinence self-efficacy questionnaire (Spek et al., 2013) asks smokers how confident they are they would not smoke in specific individual situations. This style of questioning reflects the fundamental idea in self-efficacy theory that self-efficacy judgments are made about single actions in isolation. Attempts to extend the theory to consider more prolonged behaviour were made in the development of *coping* self-efficacy (DiClemente et al., 1995; S. L. Williams, 1995), which focusses on perceived capability to maintain a behaviour in the face of adversity. However even here, notions of hierarchical action are only implicit in the theory – high-level policies of action are not explicitly discussed, nor is actual behaviour measured. Here I make an explicit claim that features of hierarchical action must be integrated with classical notions of flat self-efficacy to more completely describe how self-efficacy judgments are made. To support this claim, I found evidence for a precise sensitivity to the exponential decreases in the likelihood of successfully executing a sequence of difficult actions as the length of that sequence increase. This sensitivity requires two things: (1) accurate estimates of perceived capability to execute the difficult action once; and (2) an accurate estimate of how needing to repeat the action multiple times influences probability of overall success (i.e., a model of how well performance will accumulate). These two components capture a need to integrate classical flat notions of self-efficacy that remain important with hierarchical self-efficacy to accurate explain self-efficacy judgments of real-world behaviours.

In addition to the mechanical difficulty of executing a skilled task multiple times, I identified a second, more cognitive component of the influence the length of a behaviour has on self-efficacy judgments. In committing to a lengthy high-level course of action we are in effect committing to a restriction of all relevant choices for the duration of that course of action. For example, a commitment to vegetarianism restricts all future choices between foods to include only those that are vegetarian. I found that this restriction on future choice was sufficiently aversive to cause my participants to deviate from the optimal solution to the task, which was (for high accuracy levels) to engage with the high-level policy and to ignore a tempting and immediate but infrequent boost in low-level reward. Note that the boost did not change the optimal solution to the task, which was to engage with the high-level policy at high accuracy levels (though the boost did narrow the gap between the value of the two competing policies). I propose two possible interpretations for this effect.

First, it could be that my participants were averse to tying their hands to a high-level policy because they judged that they would be unable to maintain adherence to it in the face of a tempting low-level reward. This interpretation is grounded in coping self-efficacy (DiClemente et al., 1995; Williams, 1995), and in this light the effect can be compared directly to abstinence in addictive behaviours, where the temptation of an off-policy reward presents a real challenge for self-restraint. This interpretation is less about the restriction of choice for the sake of being able to pursue valuable off-policy outcomes *per se*, and more about anticipating a failure to behave in line with the high-level policy when presented with a difficult choice. It is not that people do not *want* to maintain adherence to a high-level policy *long-term*, it is that they perceive that they may be unable to reject tempting off-policy rewards, and so they do not engage with the high-level policy in anticipation of their own failure. The lengths of the commitments involved are relevant here. To achieve the high-level outcome of the bonus in my task or prolonged abstinence for a smoker, very many actions must be successfully executed. By contrast, to experience the low-level reward, only one action must be successfully executed. If I perceive that I am likely to fail to resist temptation at some point, then why engage with the high-level policy at all? In sum, when deciding whether to engage with a high-level policy, people judge how well they expect themselves to maintain that policy, and this judgment influences how willing they are to commit to the policy at all.

Second, my participants could have been averse to tying their hands to a high-level policy because they *wanted* to take advantage of any and all highly rewarding and immediately available outcomes that came their way. This interpretation is derived from prospect theory (Kahneman & Tversky, 2018), which, broadly speaking, states that losses feel worse than equivalent gains feel good. It may be that the prospect of losing out on a highly rewarding outcome due to a prior commitment to a high-level policy that requires ignoring that outcome feels worse than earning that outcome feels good. This asymmetry in how we process losses and gains would artificially inflate the negative value of the prospective losses incurred by the restrictions placed on choice by a candidate high-level policy of action. In effect, this would make that policy less appealing by applying an over-tuned penalty for these losses, thereby dissuading commitment to it. There are other heuristics that are relevant here, such as the peak-end rule (Do, Rupert, & Wolford, 2008), which shows that peaks in value are more salient for our decision making processes than more moderately valued but tonic outcomes. In sum, the peculiarities in how people estimate value might bias decisions to avoid any restrictions on choice where possible to limit prospective losses. Given that higher level behaviours will by definition place lengthier restrictions on choice, this would translate into a general bias to avoid tying our hands to higher-level policies of action.

The two interpretations of the effects observed here are not mutually exclusive. It could well be the case that people make judgments similar to those described by coping self-efficacy about their ability to maintain adherence to high-level courses of action *and* that they are biased to want to avoid restrictions on choice so as to avoid any prospective losses. These two may also interact for a compounding effect that severely dissuades commitment to high-level policies: if I deem myself unable to adhere to a high-level policy for long enough to experience the outcomes it aims to achieve *and* I expect to incur low-level losses from following it, I may be very unlikely to ever commit to it. Future research could aim to disentangle these two interpretations by varying the cognitive effort required to maintain a high-level policy independently of the prospective losses incurred by committing to it, and this may shed further light on how these two processes interact in decisions between actions at multiple hierarchical levels.

Future research could also investigate the relationship between the effects observed here and the severity of the contingencies involved. In my task and in the examples I have discussed (e.g., smoking), a single failure to adhere to the high-level policy is catastrophic with respect to the end goal the policy aims to achieve. However, not all behaviour follows such strict contingencies. If these were relaxed, such that some number of failures were permitted under the high-level policy without these failures resulting in a complete ruling-out of the high-level outcome, it is possible that the effects here would diminish. There is in all likelihood a balance, whereby the acceptability of failure under a high-level policy will determine how important the features of high-level action identified here are for self-efficacy judgments and choice.

To conclude, I aimed in this chapter to answer how self-efficacy might contribute to or resolve decision conflict between actions at different hierarchical levels. I identified the length of those actions as a key factor of interest for how the brain computes self-efficacy judgments over hierarchically higher-level actions. Over two experiments, I demonstrated first that the length of a course of action influences choice both due to the increased mechanical difficulty of executing lengthier sequences of action and due to the increased cognitive difficulty of adhering to lengthier sequences of action that place restrictions on choice. These findings outline two factors tied to a now extended hierarchical self-efficacy theory that contribute to resolutions of conflict in decisions between actions at different hierarchical levels.

Chapter 4

Subjective Biases:

Do we prefer to pursue outcomes at specific hierarchical levels?

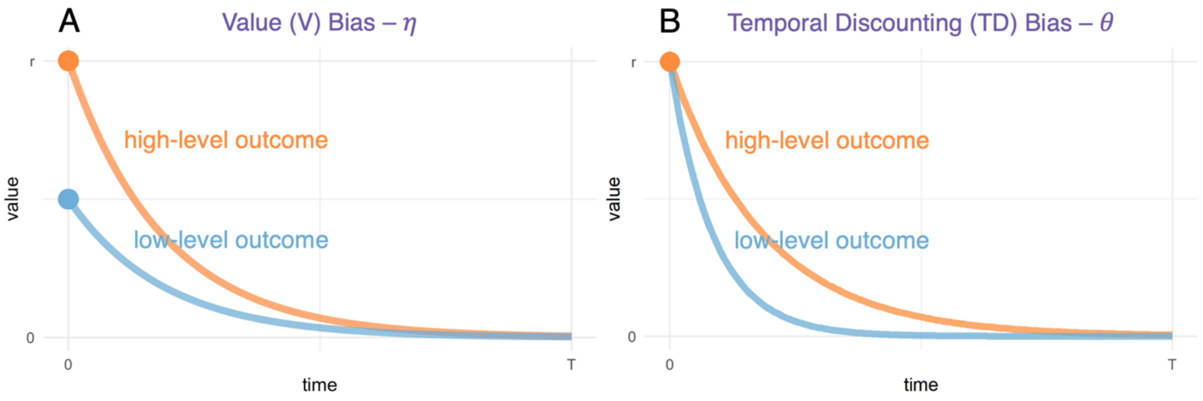
4.1 Introduction

In Chapter 2, I demonstrated that human behaviour is hierarchically organised. I then turned to focus on whether such an organisation has any influence on how we decide how to behave. In Chapter 3, I found that lengthy high-level behaviours are associated with a diminished sense of self-efficacy, and that people are less likely to commit to these behaviours as a result. In this chapter, I investigate a complementary aspect of any action decision. When deciding whether to perform an action, people consider not only the likelihood that they will succeed, but also the value of the outcomes that would follow from success. In this chapter, over three experiments I investigate whether higher-level outcomes are intrinsically and subjectively more valuable than lower-level outcomes.

Why might people be biased to prefer to pursue outcomes at specific hierarchical levels? Temporal discounting (Doyle, 2013; Kacelnik, 1997; Odum, 2011), is the a well-documented discount in the value of an outcome as it moves further away in time, is intuitively sensible. As the span of time between *now* and a prospective outcome increases, the probability of the outcome occurring decreases in an uncertain world. Is there a similarly sensible justification for discounting the value of an outcome not according to time, but according to its hierarchical level? Given that humans do discount the value of more distant outcomes (Odum, 2011) and given that higher-level outcomes are by definition further away in time, the absence of any counteracting force (i.e., a discounting of lower-level outcomes) here would make sufficiently high-level outcomes unreasonably difficult to pursue. Even outcomes high in objective value would, if at a high-enough level (and therefore far enough away in time), have their subjective value diminish to near zero thanks to temporal discounting. For example, the risks posed by climate change are very large in magnitude but may seem very far away in time, such that the negative value of the extreme outcomes involved may be discounted to near-zero thanks to their extreme distance. An inability to act on high-level outcomes with high-level courses of action is ultimately maladaptive – in the case of climate change, this inability would lead to a species-level existential risk being unanswered. However, this could be rectified should the human brain discount value differently at different levels of the behavioural hierarchy.

There are two ways in which hierarchical level-dependent biases in subjective value could be implemented (see Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s, a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward.). First, it could be that there is a global boost in subjective value for higher-level outcomes. This would elevate the value of any high-level outcome, though it would still be subject to the same rate of temporal discounting as a low-level outcome, and so this boosted value would still decay quickly over time. Second, it could be that the rate of temporal discounting applied to any outcome is dependent upon the hierarchical level of that outcome. This would not alter the subjective value of two comparable high and low-level outcomes, but it would allow for high-level outcomes to maintain their value over a longer span of time. This implementation more directly addresses the issue outlined with distant but highly valued high-level outcomes being ignored, as higher-level outcomes could be subject to slower temporal discounting than lower-level outcomes.

Over three experiments, I test whether humans hold any such level-dependent subjective bias. My approach was to first test for whether any such bias was present in human behaviour, and if so, I intended to then investigate which of the two mechanisms proposed in Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s,a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward. was responsible for this bias. In experiment 4-1, I pit two policies of action that earn equal sums of reward but different distributions of reward over high and low levels against one another. Consistent with the absence of any level-dependent bias, I find no general preference in the population to prefer to pursue high- (or low-) level outcomes. Note that for any bias of this sort to exist, I would expect it to be consistent within the population, with any inter-individual differences manifesting only in the strength of the bias and not in its direction (as is the case for temporal discounting, Odum, 2011). In two follow-up experiments, I investigate whether any subjective bias in the evaluation of high-/low-level outcomes



$$Q(\mathrm{s,a}) = \eta^{\mathrm{H}} \cdot (\alpha \cdot (\theta^{H} \cdot \gamma^{t}) \cdot r) - Q(\mathrm{s,a})$$

Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s,a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward.

can be produced by context. I introduce a social-attentional cue to the same task used in experiment 4-1, and I find that preference can indeed be shifted to favour high- or low-level outcomes as per this social-attentional cue. While people seem to hold no general preference to pursue high-level outcomes, it seems that such preferences can be induced by even minimal social cues to guide attention.

4.2 Experiment 4-1

4.2.1 Methods

4.2.1.1 Participants

This experiment was approved by the UCL Research Ethics Committee, and it was hosted online. Participants were recruited via Prolific (www.prolific.co), and all were then redirected to a personal website where the experiment was hosted. All participants provided informed consent prior to the start of the experiment. As in the previous chapters, a sequential analysis approach was used to ensure the study was sufficiently powered while avoiding Type 1 errors. Sequential analysis involves collecting and analysing data at increasingly large sample sizes while controlling for the Type 1 error rate by lowering the threshold for significance (see Lakens, 2014). I planned to collect data in batches of 40 subjects, to analyse the data, and to then decide whether results were convincing enough to conclude data collection. If no conclusive results were found, I would then continue data collection, but I would halve my threshold p-value for the next iteration. Note that a post-tutorial questionnaire was included and failure to answer any of the questions here resulted in early termination of the experiment (though I continued data collection until a full sample of 40 participants had fully completed the experiment). From my first sample of 40 participants, I found no evidence for any effect of hierarchical level on choice and no indication that such an effect might be present in human behaviour, and so I terminated my sequential analysis plan at the first step. I therefore had a final sample size of 40 participants for this experiment. Participants received £5 per hour as a base rate and could earn an additional bonus payment (up to £3) based on performance in the task.

4.2.1.2 Design

The experiment consisted of six blocks of five trials of a delivery task, with each trial requiring that the participants select and deliver a single package. Blocks in the task corresponded to weeks, and trials to days of the week (Mon-Fri). Each trial presented a choice between one of two possible deliveries (as marked by orange and blue packages), the task being to select a package and then deliver it at its destination. The rewards were structured such that one of the two packages would earn a lesser amount per individual delivery (6 points vs 10) but could earn a bonus amount that would make up for this difference if repeated over all five trials within a block (the assignment of orange or blue to either reward scheme was randomised over participants). I refer to the package that could earn the high-level bonus as the high-level package and the package that maximised low-level payments for individual deliveries as the low-level package.

The reward structure used in the task presented participants with a choice between two competing policies of action. First, selecting the low-level package on all trials in a block would maximise low-level reward and would earn a total of 50 points (10 points per package). Second, selecting the high-level package on all trials in a block would earn less low-level reward (30 total; 6 per package) but would secure a high-level reward in the form of a bonus (20 points). I refer to the former as the *low-level policy*, and the latter as the *high-level policy*. Critically, the bonus payment occupies a higher hierarchical level than do the payments earned for individual deliveries. The bonus is contingent upon a policy of action taken over an entire block, whereas payments for individual deliveries depend only upon behaviour within a single trial. Therefore, although the two policies earn an equal objective sum of reward, they differ in the distribution of that reward over the behavioural hierarchy. Where the low-level policy earns only low-level reward, the high-level policy earns a mixture of high- and low-level rewards. As a result, if my participants held a general preference for pursuing high- over low-level outcomes, I would expect them to prefer the high-level policy over the low-level policy.

To rule out any effect of temporal discounting, all reward was delivered at the end of the block. This allowed a contrast of high- vs low-level policies with the only differentiating factor being the distribution of rewards over high and low levels of behaviour. If I did detect a preference for one policy over another, I could therefore attribute this to a genuine preference for higher- or lower-level outcomes. Note however that this design decision does make a disentanglement of the two potential biases outlined in Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s, a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward. impossible – I could detect a bias, but I could not attribute any bias conclusively to either level-dependent temporal discounting rates nor to a level-dependent global boost in value.

Comparisons between optimal policies of action make sense only if my participants know what those policies are with certainty. If a participant did not know how much reward was available under the high- and low-level policies, then they could not make an informed decision between them. To verify that all participants had complete knowledge of the rewards available under each policy, participants completed an extensive tutorial that would present all information necessary. The tutorial ended with a questionnaire which probed for complete understanding of the rewards available under each policy and an understanding that the policies earned the same objective sum of reward. To proceed on to the main experiment, all questions here must have been answered correctly, and so all participants in my data held complete knowledge of the reward structure, allowing for a clear interpretation of their choices in the task.

4.2.1.3 Procedure

The task was organised into six blocks of five trials, and as mentioned payment for the block was provided only after all trials were completed. Prior to the task itself, participants completed a tutorial that explained the task and the reward structures in detail (see Figure 4-2 – Illustration of procedure followed by experiments 1, 2, and 3. for a full summary of the procedure followed).

Tutorial

The tutorial began with an introduction to the environment within which participants were to be making deliveries (this comprised a 10x10 grid with some basic obstacles restricting movement). Participants were then given an opportunity to navigate around the environment by moving their avatar up, down, right, and left using the corresponding arrow keys. Once they were comfortable with how to move around, the participants were asked to complete a series of practice trials. The first two practice trials had the participant select and deliver an orange package before doing the same for a blue package (the order of the colours was randomised over participants).

Blocks in the task corresponded to weeks, and participants were to complete one delivery per day Monday to Friday (corresponding to 5 trials per block). Participants were able to track their progress through any given week by looking to the top left of the on-screen display, where they could find a record of the days of the week with all completed days coloured in according to the package delivered on that day. A third and final practice trial was set up as the fifth and final trial of an in-progress block to introduce this to participants. On this trial, participants were free to decide for themselves between the two packages. Participants then received

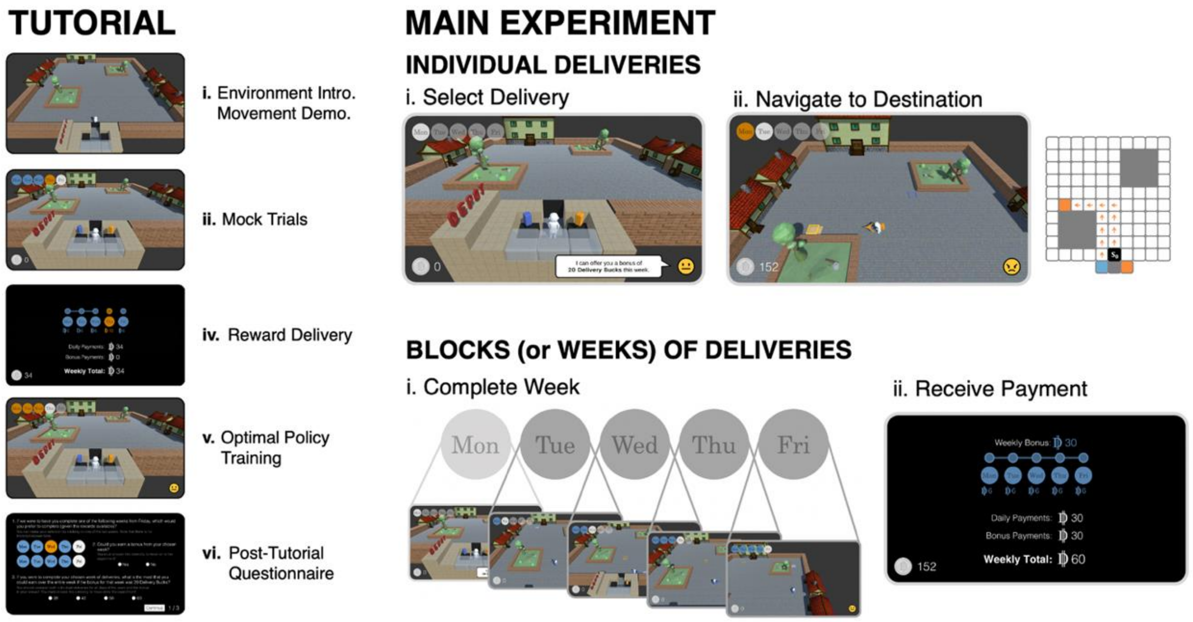


Figure 4-2 – Illustration of procedure followed by experiments 1, 2, and 3.

payment for the week as though they had completed it all themselves. This introduced the idea that individual payments earned for delivering orange or blue packages were unequal (with one of the two packages being more valuable than the other), and that there was a bonus available each week. Participants were told that they could earn the bonus payment each week should all deliveries within a week be of the less rewarding package.

Participants were then trained on the two competing policies that exist within the task (see Design). This was achieved by having participants complete two in-progress weeks (with Mon-Wed deliveries already completed) repeatedly until they earned as much as possible given the deliveries already completed. One of the two weeks required that participants select the high-level package every day to earn the bonus; the other required that participants ignored the bonus and selected only the low-level package.

To end the tutorial, participants answered three sets of questions on the content they had covered. Each set consisted of the same three questions: (1) participants were to select which of two in-progress weeks they would most like to complete given the rewards available to them for doing so; (2) for their selected in-progress week, participants were to judge whether the bonus was attainable; (3) again for their selected in-progress week, participants were to estimate the maximum amount they could earn by completing the week. These questions were chosen to ensure that participants (1) understood how to attain the bonus, (2) understood that the two competing policies offered greater earnings than any other sub-optimal alternative, and (3) understood that the two competing policies offered objectively equal pay-outs for particular values of the bonus (i.e., on neutral blocks). Importantly, should participants fail to answer either of the final two questions correctly for any of the three sets, they were excluded from the remainder of the experiment. The first two sets of questions pitted each of the two competing policies in turn against sub-optimal alternatives, where the third set posed participants with a choice between the two competing policies.

Main Experiment

The experiment was organised into six blocks of five trials. Each trial began with a selection between an orange and a blue package (the position of the packages on the right or left of the screen was randomised over all trials). Having selected a package, an avatar would automatically leave the depot and the destination would be revealed (this was concealed during selection so as not to bias decisions between packages). Participants would then navigate to the destination, and upon entering the appropriate tile on the grid the package would be delivered and the next trial would begin. At the end of each block, participants received payments for the individual deliveries made and any bonus they may have earned and then the next block would begin. Once all six blocks were completed, participants were required to answer a closing question that asked for a justification of their preferred policy ("Please tell us whether (in general) you preferred to go for the bonus or ignore it and why?").

4.2.1.4 Analysis

Our analysis plan for the current task was simply to investigate whether the average preference for each participant (computed as the proportion of times a participant selected the high-level policy over all six blocks) deviated from the chance level of 0.5. I planned to compute preference for each participant, and to then perform a one-sample t-test against the chance level to detect any significant deviation.

4.2.2 Results

I found no evidence to suggest that people have a general bias to prefer to pursue high- or low-level outcomes. Analysis of the proportions of blocks completed by my 40 participants under the high-level policy revealed no significant deviation from the chance level of 0.5 ($t(39) = -0.66, p = .511$). The distribution of proportions does appear to be tri-modal, with some participants opting to commit to the low- or high-level policies on all blocks in the experiment, and others hovering around the chance level of 0.5 (see Figure 4-3 – Smoothed density of proportions of blocks completed under the high-level policy for all participants. For each participant, I computed the proportion of blocks completed under the high-level policy, and I show here the density of those proportions over the full range.). Whilst this may well indicate that there are inter-individual differences in preference for high- vs low-level outcomes, it may also be a result of the fact that in the absence of any bias to nudge preference in the direction of either outcome, participants are free to select any strategy they like to complete the blocks under one or other of the optimal policies. The lack of any general shift in the population in one direction or another indicates that people are not biased in how they evaluate outcomes at different hierarchical levels.

4.2.3 Discussion

I found no evidence to suggest that humans are biased in how they estimate the value of outcomes that occupy different levels of a behavioural hierarchy. The lack of any consistent shift in my population of participants in favour of high- or low-level

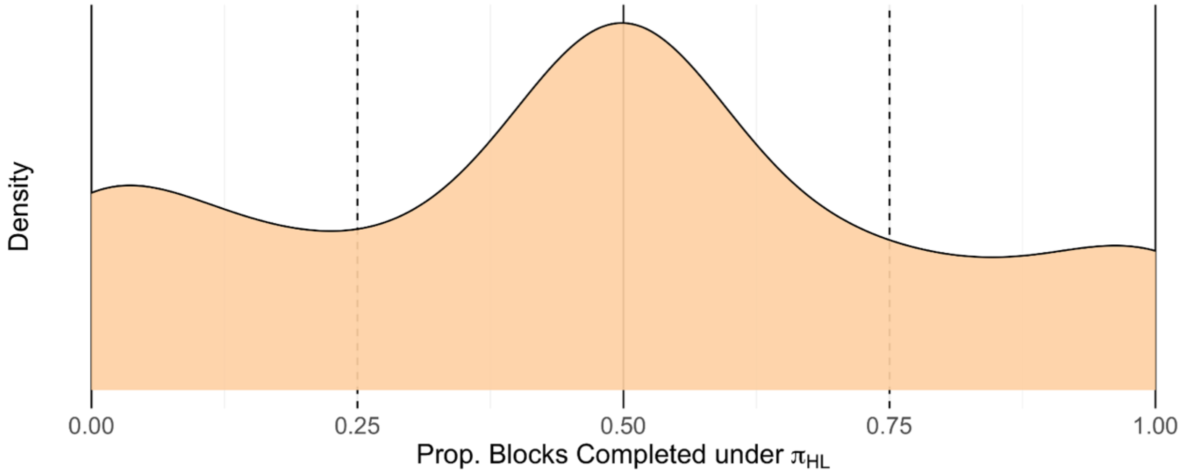


Figure 4-3 – Smoothed density of proportions of blocks completed under the high-level policy for all participants. For each participant, I computed the proportion of blocks completed under the high-level policy, and I show here the density of those proportions over the full range.

outcomes does not support the concept of level-dependent biases. Failure to relax the severity of temporal discounting for high-level outcomes (by way of some subjective bias) makes it difficult to see how any individual motivates action in their pursuit. Yet, people are in fact able to motivate actions directed at high-level outcomes (for one example, see prevalence rates of vegetarianism: Paslakis et al., 2020). We can rule out any general and global bias to prefer high level outcomes on the basis of my current experiment, but there may well be other factors that contribute to high-level preferences than hierarchical level alone.

The high-level outcomes of particular interest here are those that are high in magnitude, making them valuable to pursue, but very high-level, meaning that their value decays to near-zero given their distance in time. The claim that temporal discounting is maladaptive here rests on these two requirements: the claim is that due to extreme distance, extremely rewarding/punishing outcomes that people *should* pursue/avoid are ignored. The outcomes involved in climate change are a good example: the cost of failing to act is immense, and yet the distance to those outcomes makes motivating action difficult. However, note that at these very high levels of behaviour, the quality of the goal shifts from being one that can be feasibly attained by an individual person to one that requires input from many people. This is a straightforward extension of a hierarchical organisation of behaviour: just as individual actions can be sequenced, the actions of individual people can be combined. We see this in issues like climate change: the action of any individual is unlikely to dent the problem, but the joint action of many people will have impact. The outcomes and the behaviours that achieve them therefore take a different shape. Vegetarianism, as one example, has become a social identity (Nezlek & Forestell, 2020), and research into the social psychology of vegetarianism reveals that motivation for and maintenance of vegetarianism are tightly linked to one's social context (for a review, see Rosenfeld, 2018). Such psychological factors are interesting but are scarcely captured by the relatively simple experiment reported here. While experiment 4-1 found no direct evidence to support the presence of a general bias to pursue high- over low-level outcomes (or vice versa), the outcomes involved in the task did not quite match the nature of the high-level outcomes identified as problematic, because they included no social context. As a result, we do not know what effect social context has on preference for high- vs low-level outcomes, which is relevant for the high-level outcomes discussed. Effectively, real-world choices might be more successful in triggering high-level choices than the simple scenarios described here.

In experiment 4-1, I presented participants with a choice between two policies of action that differed only in the distribution of reward over two hierarchical levels, and I found no population-level preference to pursue high-level outcomes over low-level outcomes making the prospect of any general subjective bias here unlikely. Given that high-level outcomes of the sort I am interested in here tend to involve group action towards a common good, I hypothesised that introducing a social context might allow us to manipulate preferences to pursue outcomes at specific levels. Therefore, over two experiments I introduced social cues to the task used in experiment 4-1 to draw attention to one or other of the two hierarchical levels included in the task. I found that introducing even minimal social cues systematically shifted preference in favour of the cued level, supporting the idea that although there is no general bias to prefer to pursue outcomes at higher hierarchical levels, a preference can be manufactured by a social context that guides attention.

4.3 Experiments 4-2 & 4-3

4.3.1 Methods

4.3.1.1 Participants

Both experiments 4-2 and 4-3 were approved by the UCL Research Ethics Committee, and both were hosted online. Participants were recruited via Prolific (www.prolific.co), and all were then redirected to a personal website where the experiment was hosted. All participants provided informed consent prior to the start of the experiment. As for experiment 4-1, a sequential analysis approach was used to ensure the study was sufficiently powered while avoiding Type 1 errors (Lakens, 2014). I planned to collect data in batches of 30 subjects (for each experiment), to analyse the data, and to then decide whether results were convincing enough to conclude data collection. If no conclusive results were found, I would then continue data collection, but I would halve my threshold p-value for the next iteration. Note that a post-tutorial questionnaire was again included and failure to answer any of the questions here resulted in early termination of the experiments (though I continued data collection until a full sample of 30 participants had completed the experiment). From my first samples of 30 participants, I found conclusive and statistically significant results, and so I terminated my sequential analysis plan at the first step. I therefore had a final sample size of 30 participants for each of experiments 4-2 and 4-3. Participants received £5 per hour as a base rate and could earn an additional bonus payment (up to £3) based on performance in the task.

4.3.1.2 Design & Procedure

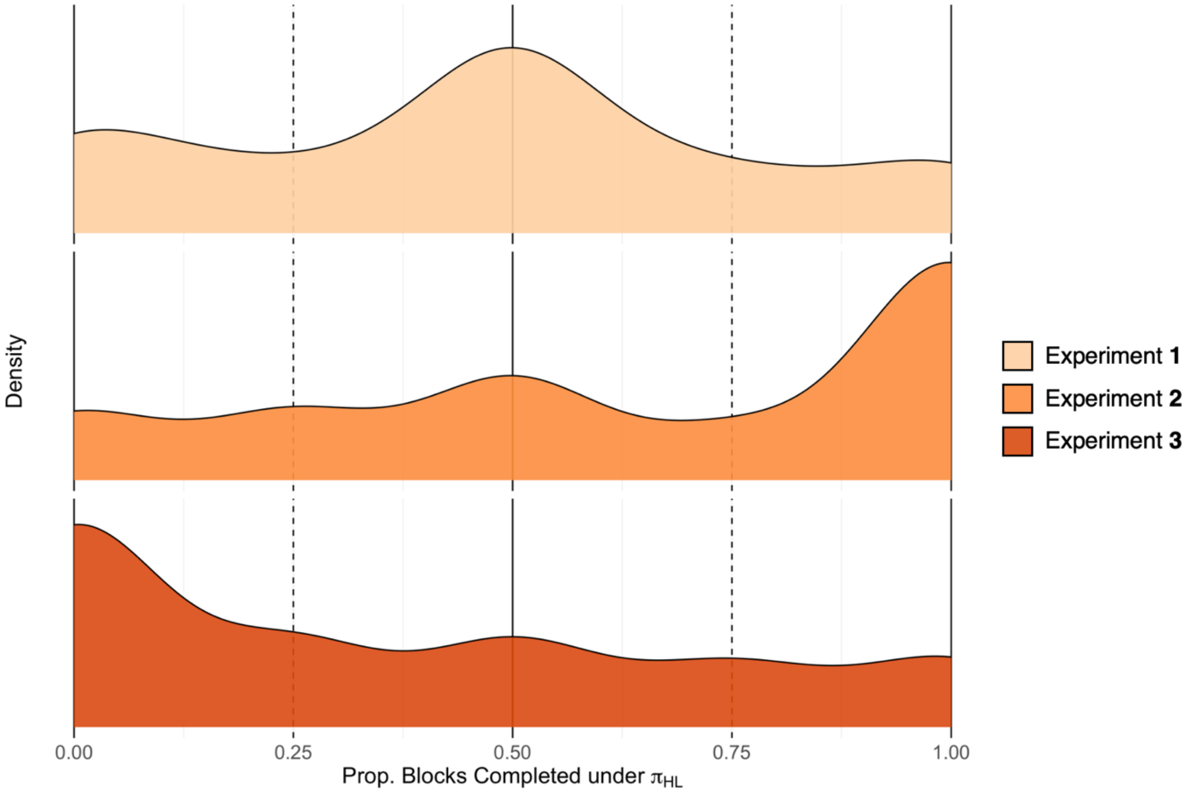
Experiments 4-2 and 4-3 followed the same design and procedure as experiment 4-1 (see Figure 4-2 – Illustration of procedure followed by experiments 1, 2, and 3.) but for two changes. First, the amount of reward available in the bonus and in the low-level payments earned by the low-level package could vary. Second, a social-attentional cue was introduced to the task in the shape of a “boss” character that would instruct the participants as to what rewards were available for each block. In experiment 4-2, the high-level bonus could change between blocks, but low-level rewards were fixed, while in experiment 4-3 the low-level rewards earned by the low-level package could change between blocks, but the high-level bonus was fixed. These changes in reward were set so that on four of the six blocks the high-level and low-level policies were of equal objective value as these blocks followed the same reward scheme used in experiment 4-1. Of the two remaining blocks, one would make the high-level policy more rewarding (by increasing the bonus in experiment 4-2 and by decreasing the low-level payments in experiment 4-3), and one would make the low-level policy more rewarding (by decreasing the bonus in experiment 4-2 and by increasing low-level payments in experiment 4-3). On each trial, the “boss” would appear at the bottom right of the screen and would provide exact information of the rewards available block by block (in experiment 4-2, the cue read “I can offer you a bonus of X this week”; in experiment 4-3, the cue read “I can offer you payments of X this week”). I made these changes with two goals: (1) adjusting the relative values of the high- and low-level policies allowed us to measure whether my participants remained sensitive to the rewards involved in the task; and (2) introducing a “boss” figure added a minimal social cue that would draw the attention of my participants to one or other of the policies (high-level in experiment 4-2 and low-level in experiment

4-3). I hypothesised that this minimal social cue would shift preference in favour of the cued outcomes and their associated policies.

4.3.1.3 Analysis

Our analysis plan for experiments 4-2 and 4-3 was similar to that used for experiment 4-1 – I would compute the proportion of blocks completed under the high-level policy as a measure of preference, and I would test (with one-sided one-sample t-tests) for a deviation in this population of preferences from the chance level of 0.5. I had clear predictions here, which were that deviations would be in the direction of the outcome cued by the social cues introduced to the two experiments, i.e., I would find a shift in favour of the high-level policy in experiment 4-2 and a shift in favour of the low-level policy in experiment 4-3. To evaluate my original hypothesis for this chapter, which was that there may be subjective biases in how outcomes at specific hierarchical levels are evaluated, I also fit reinforcement learning models to behaviour. For all possible proportions of blocks completed under the high-level policy, I fit three RL models: one was a simple Rescorla-Wagner model, and the other two extended on this simple baseline by implementing one or other of two biases outlined in Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s, a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward.. These biases capture (1) level-dependent rates of temporal discounting, and (2) a global boost in value for high-level outcomes. I fit these three models to preference data, and I compared their fit using the Bayesian information criterion, which would penalise the latter two models for their additional parameter (which controlled the strength of their respective biases). The intention here was to see whether introducing a bias led to significant improvement in the fit of my models to data, which would, if true, support the idea that such biases exist and can be prompted by social context.

4.3.2 Results



Density
0.00
0.25
0.50
0.75
1.00
Prop. Blocks Completed under π_{HL}
Experiment **1**
Experiment **2**
Experiment **3**

When cued to attend to outcomes at specific hierarchical levels, my participants

Figure 4-4 – Smoothed density of proportions of blocks completed under the high-level policy for all participants in each experiment. For each participant, I computed the proportion of blocks completed under the high-level policy, and I show here the density of those proportions over the full range.

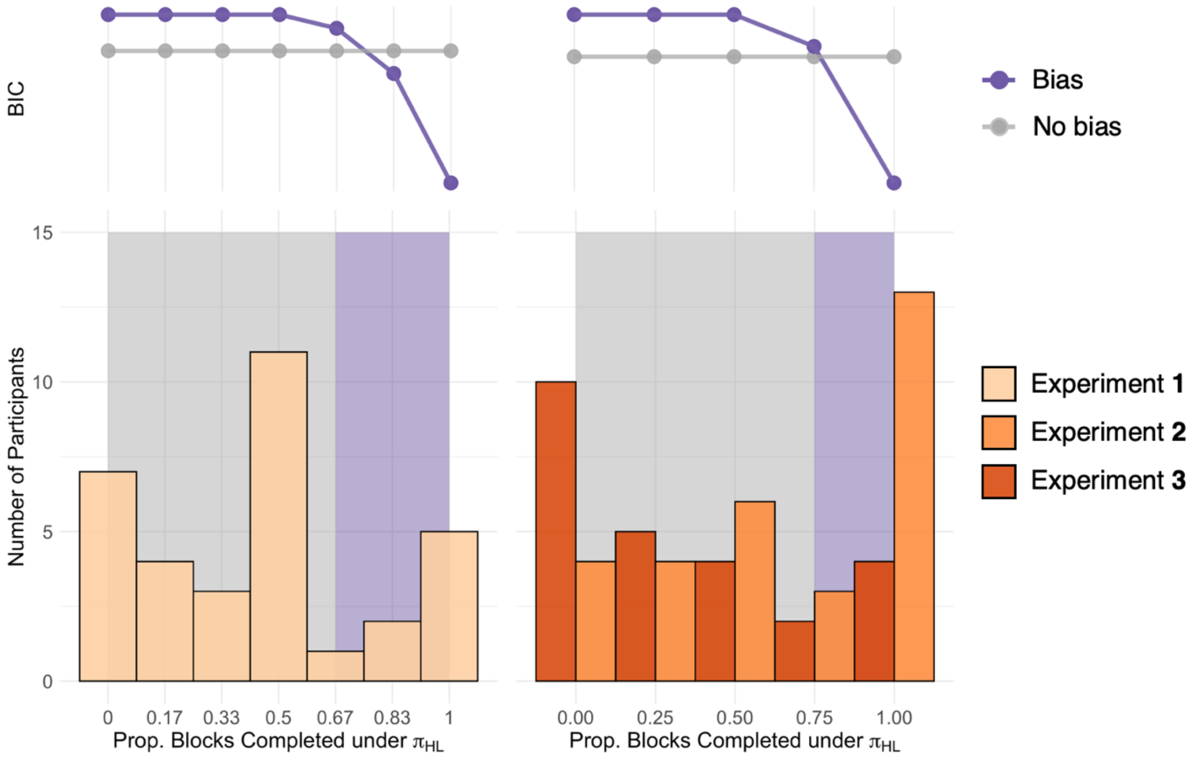


Figure 4-5 – Fits of bias and no-bias models to the full range of proportions of blocks completed under the high-level policy. The top panel plots BIC values for the two models for comparison of fit (lower BIC = better fit). The bottom panel plots histograms of proportions for experiment 4-1, which included no social cues, and experiments 4-2 & 4-3, which included social cues to pursue the high-level and low-level outcomes respectively.

tended to pursue those cued outcomes (see Figure 4-3 – Smoothed density of proportions of blocks completed under the high-level policy for all participants. For each participant, I computed the proportion of blocks completed under the high-level policy, and I show here the density of those proportions over the full range.). Where there was a social cue to attend to the high-level outcome (in experiment 4-2) there was a significant shift in preference in favour of the high-level policy $(t(29) = 2.07, p = .048, d = 0.37)$, and where there was a social cue to attend to the low-level outcome (in experiment 4-3) there was a significant shift in preference in favour of the low-level policy $(t(29) = 2.10, p = .045, d = -0.38)$. It seems therefore that minimal social cues are sufficient to prompt a bias in preference to pursue outcomes at specific hierarchical levels.

For a full comparison of behaviour on all three experiments, I fitted RL models that did or did not include a bias to pursue specific hierarchical levels to all possible levels of preference across all three experiments. The no bias model implemented classic temporal difference Q-learning (as in the equation in Figure 4-1 but without either of the two biasing parameters θ and η). The two bias models implemented each of the two biases outlined in Figure 4-1. Note that although both models were implemented, they were equivalent in terms of performance. Learning rates for all models were fixed at a value of 1 to reflect the fact that human participants were overtrained on the task and had perfect knowledge of the competing values of high- and low-level policies. We thus fit the biasing parameters and the temporal discounting rate as free parameters to the behavioural data using maximum likelihood estimation.

I compare the fits of these models using the BIC, which penalises the bias models for their inclusion of a bias parameter (in addition to the base learning rate and temperature parameters included in all models). Note that the fits of the two alternative bias models (TD and V) are equivalent given that the strength of the bias can be equivalently adjusted in both models by adjusting their respective bias parameters. I find that only at very high levels of subjective bias do the bias models provide a better account of behaviour (see Figure 4-5 – Fits of bias and no-bias models to the full range of proportions of blocks completed under the high-level policy. The top panel plots BIC values for the two models for comparison of fit (lower BIC = better fit). The bottom panel plots histograms of proportions for experiment 4-1, which included no social cues, and experiments 4-2 & 4-3, which included social cues to pursue the high-level and low-level outcomes respectively.). These high levels of bias were only found consistently in experiments 4-2 and 4-3, meaning that the simpler no-bias model is the best fit to behaviour in experiment 4-1 and the bias models fit best to behaviour in experiments 4-2 and 4-3. Note also that my models assume that any subjective bias in the evaluation of outcomes at specific hierarchical levels is in favour of high-level outcomes, i.e., a bias here inflates the value of higher-level outcomes. As such, my bias models do not capture behaviour in experiment 4-3, in which most participants prefer to pursue the low-level outcome in line with the social cues presented to them. A relaxation of my assumption that subjective biases are always in favour of higher hierarchical levels however would result in a symmetrically good fit for low-level preference. Taken together, the performance of my models demonstrate that (1) in the absence of any social cues to guide attention, there is no general bias in favour of high-level outcomes (see experiment 4-1); (2) biases to pursue outcomes at specific hierarchical levels can be readily induced by social cues (see experiments 4-2 & 4-3); and (3) biases in how people evaluate outcomes at specific hierarchical levels do not always inflate the value of higher-level outcomes (as in experiment 4-2): they can also shift preferences in favour of lower-level outcomes (as in experiment 4-3).

4.3.3 Discussion

Over two experiments I demonstrated that introducing a minimal social cue to attend to outcomes at specific hierarchical levels can shift preference in favour of whichever level is cued. In experiment 4-2, I found that a cue to attend to a high-level bonus delivered by a *boss* figure would bias choice to pursue the bonus. In experiment 4-3, the cue was instead for low-level payments and here choices were biased to maximise low-level reward. I fit simple RL models to choice behaviour, and found that introducing a level-dependent bias in how outcomes are evaluated provided a marked improvement in the fit of these models to participants that were extreme in their preferences (i.e., participants that would select high- or low-level policies on all blocks). Taken together, these results indicate that a preference to pursue outcomes at specific hierarchical levels can be manufactured by social context.

What was it about the social cue introduced to experiments 4-2 and 4-3 that prompted a shift in preference? One interpretation is that the participants engaged with a learning process similar to that described by social reinforcement leaning, which extends standard reinforcement learning to include the many social interactions that might be relevant for learning how to behave (Isbell, Shelton, Kearns, Singh, & Stone, 2001; Jones et al., 2011; Najar, Bonnet, Bahrami, & Palminteri, 2020). Najar et al. (2020) investigated precisely how social signals influence human learning by comparing three plausible algorithms for the influence of social signals on learning: (1) *decision biasing* postulated that imitation consisted of a transient bias to select action in line with an agent's peers without affecting the agent's value function; (2) *model-based imitation* postulated that the agent would infer the value function of a peer and use that inferred value function to guide action without replacing their own; and (3) *value shaping* postulated that the agent would update their own value function according to the actions of a peer. They compared these three alternative mechanisms in two novel social reinforcement learning tasks. Model comparison revealed that value shaping provided the best fit to human behaviour. Value shaping, that is the updating of one’s own value function according to the actions of one’s peers, lines up neatly with my proposal that participants were biased in how they estimated the value of outcomes at specific levels according to a social cue to attend to those outcomes. In both cases, there is a social signal that some action or outcome is somehow important, and in both cases this social signal directly influences computation of value.

A second, alternative interpretation of the results discussed is that participants were biased to pursue cued outcomes simply because they were more salient. The cue would draw attention to one or other of the two policies, and the salience of that policy might then have biased choice in its favour. Indeed, people do tend to place disproportionately high weight to more salient outcomes (in e.g., consumer choice: Bordalo, Gennaioli, & Shleifer, 2013). However, the outcomes involved in my tasks are of equal objective value, and participants knew this with certainty. It is therefore unclear how salience would bias *preference*. In consumer choice, for example, salience exerts an influence by drawing attention to one feature (such as price) over others (such as quality) such that decisions are based on the more salient features. However, alternative choices still vary along these more salient dimensions. In my case, there is no variability along the major decision-relevant dimension (points earned) between the two alternative policies. Thus, although the cues involved in the task did make one or other of the two policies more salient, all that is really drawn to my participants’ attention is that the two policies are of equal objective value. Attention can be used to collect information that can break symmetrically valuable options (Brunton, Botvinick, & Brody, 2013), but attention alone does not break equalities.

I favour the first of these two interpretations: that biases in preference here are due to a process of social reinforcement learning, where the social signal to attend to outcomes at a specific hierarchical level influences computations of value. I favour this interpretation for two reasons. First, there is correspondence between the behaviour I observed here and demonstrations of value shaping in social contexts (Najar et al., 2020). Second, alternative explanations that appeal to attention and salience do not easily discriminate between the two candidate policies of action here, given that they are of equal objective value. Therefore, I suggest that social context can bias the evaluation of outcomes that fall at specific hierarchical levels, introducing a subjective and level-dependent bias in people's preferences for pursuing high- or low-level outcomes.

4.4 General Discussion

In this chapter, I aimed to investigate whether people are biased in how they estimate the value of outcomes at specific hierarchical. I started with the hypothesis that it would be rational to bias evaluation in favour of higher-level outcomes given that these outcomes are necessarily further away in time and are therefore subject to more severe temporal discounting. However, in experiment 4-1, I found no evidence for such a bias. There was no indication of even a slight shift in favour of either high or low-level outcomes, suggesting that there is no hierarchical level-dependent bias in how humans evaluate outcomes. In two follow-up experiments, I investigated whether such a bias, if not present in a vacuum, could be induced by social context. High-level outcomes of the type identified here (i.e., those that are high in magnitude but far away in time) tend to require group action, and so I hypothesised that introducing social cues to attend to outcomes at specific levels might produce a preference for those outcomes. Consistent with this hypothesis, I found that minimal social cues were sufficient to produce a preference for outcomes at either of the two hierarchical levels included in the task. Taken together, these results suggest that although people hold no intrinsic, general bias to prefer to pursue high-level outcomes, such a bias can be produced by social context.

Is it sub-optimal to hold no general bias to value high-level outcomes more highly than low-level outcomes? Answering this question depends on our definition of optimal. We can define the optimal course of action as the one that would maximise expected long-run sum of reward. However, this long-run sum is complicated by temporal discounting (Odum, 2011), which will discount the value of a reward as it moves further away in time. The complication arises due to the fact that higher level outcomes will necessarily be further away in time, and so they will always be subject to more severe temporal discounting (than would a lower-level reward). This may well be a reasonable imbalance: higher-level outcomes require lengthier spans of action to achieve them, which requires a larger commitment and is associated with a lesser certainty of success. However, for some high-level outcomes that are very large in magnitude, this imbalance become more a barrier to normative action than a reasonable conservatism around making lengthy and uncertain commitments. Consider climate change, or one's personal health. Outcomes under both examples are far away in time but potentially immense in magnitude (see also *Pascal's wager*; Connor, 2007). Is it reasonable to devalue the destruction of the planet's ecosystems or the degradation of our own health and wellbeing to zero purely because these outcomes are distant, and we have no counteracting bias to allow an adaptive pursuit of the high-level outcomes that matter? This is an open and philosophical question, but it is one that is important to ask in light of my finding that there appears to be no subjective bias to value high-level outcomes more highly.

Rather than finding a general bias to value high-level outcomes more highly than low-level outcomes, I found that I was able to influence preferences by introducing social cues to attend to specific outcomes at specific hierarchical levels. This is a distinctly different bias to the one I searched for in experiment 4-1 – I searched for a general bias to prefer to pursue *all* outcomes that occupy specific hierarchical levels, but what I found in experiments 4-2 and 4-3 was that preference for one outcome or another was influenced by social cues. Note that this is not a level-dependent bias per se, but a bias to pursue a specific outcome that does itself occupy a specific hierarchical level, and a bias of this sort is in line with the social reinforcement learning literature in which social influences on choice are best captured by biases in how value is computed (Najar et al., 2020). Although the bias here was not itself purely hierarchical, social cues could interact with a hierarchical organisation of behaviour in important ways.

There are two reasons that might underpin a social cue to attend to a specific outcome: (1) a peer/group might be sharing their preference (e.g., "You should eat the burger because it tastes good."); or (2) a peer/group might be sharing a common goal (e.g., "You should not eat the burger because eating meat contributes to climate change."). These two cues differ in the hierarchical level on which they operate: the former is a lower-level cue, as it provides information that bears on a single choice, whereas the latter is a higher-level cue, as it encourages the individual to act in line with a shared policy of action towards some shared high-level goal. Future research could investigate whether and how these distinct cues of shared preferences or common goals contribute to decisions between actions directed at high- and low-level actions to elaborate further on how social context interacts with hierarchy.

Sharing a common goal makes a connection between the way any individual person organises their behaviour and the organisation of a group of people. Individual people organise their behaviour hierarchically, and at sufficiently high levels a natural progression of this hierarchical organisation is to move beyond the action of the individual alone and towards a grouping together of the actions of multiple people. This provides a natural way for social cues to be integrated with hierarchically organised behaviour; a cue that one person is working towards a goal that requires collaboration offers another person the possibility to collaborate with them towards that goal. Social influences on choice therefore describe a straightforward extension of hierarchically organised behaviour to higher-level outcomes that require collaboration.

The social context dependent bias observed here goes some way to resolving the problems that occur when temporal discounting for high-level outcomes reduces their effective value. However, the mechanism by which this social bias operates remains unclear, and I can provide no precise account of how the bias operates. I speculate that the observed bias may be similar in function to the value shaping algorithm put forward by Najar et al. (2020), which adjusts the value function held by an individual according to social information. As an extension of this, I proposed two possible methods by which a subjective bias in the evaluation of outcomes at different hierarchical levels could be implemented: (1) first, I suggested that temporal discounting could be eased for higher-level outcomes allowing their value to persist over time; and (2) second, I suggested that a global boost in value could be applied to progressively higher-level outcomes (see Figure 4-1 – Two possible implementations for a subjective bias in the evaluation of outcomes at specific hierarchical levels: (A) a global boost in value (implemented here by η) could be applied to outcomes at specific hierarchical levels, even though both outcomes share a common discounting function; and (B) different temporal discounting rates (implemented here by θ) could be applied for outcomes at different hierarchical levels, even though the outcomes share a common *current* value. Other parameters are as follows: $Q(s, a)$ Q value associated with the state action pair (s, a); α – learning rate; γ – temporal discounting rate; r – reward.). The former suggestion would, I argue, have been the better description of any general preference for high-level outcomes given that the rationale for this preference would be to counteract the effect of temporal discounting for necessarily distant outcomes. However, the latter of these two suggestions aligns better with value shaping. It is intuitive for a social cue to act towards a particular outcome by boosting the value of that outcome as means of demonstrating the social approval tied to it. For example, celebrity endorsement of a product boosts the value of that product *now*, and it does not merely diminish the rate at which the product loses its value with time. I cannot conclusively disentangle these two possibilities, however, as each possibility offers an equally good description of my data. Future experiments could resolve this by varying the time at which equally valuable outcomes that occupy distinct hierarchical levels are delivered.

To conclude, I found no general preference to pursue high-level over low-level outcomes. I argue that this can lead to sub-optimal behaviour for particularly high-level and particularly valuable outcomes that suffer a decay in their value which makes motivating action in their pursuit difficult. However, this sub-optimal behaviour is resolved by social context, which can prompt a bias to pursue specific outcomes. I suggest that this social context dependent bias can be interpreted as a natural extension of hierarchically organised behaviour to a level above the action of any individual and into the joint action of a group. A complete understanding of how exactly this bias functions and how it interacts with hierarchically organised behaviour is important, as many of the most pressing real-world problems fall into the category of those outcomes discussed here. For two examples, the outcomes involved in climate change and in personal health are distant in time and require lengthy policies of action to fulfil them. Understanding how people counteract the influence of temporal discounting on these outcomes given their high-level nature is an important component of understanding adaptive human behaviour in the real world.

Chapter 5

General Discussion

5.1 Summary

This thesis aimed to investigate how a hierarchical organisation of human action might influence how people decide to act. The hierarchical nature of human action is well evidenced (Barto, Konidaris, & Vigorito, 2014; Botvinick, 2008; Conway & Christiansen, 2001; Diedrichsen & Kriegeskorte, 2017; Koechlin & Summerfield, 2007; Lashley, 1951; Rhodes et al., 2004a; Sakai et al., 2003). Most descriptions of this hierarchical structure appeal to a sequencing of lower-level actions to produce higher-level routines of behaviour (Lashley, 1951; Sakai et al., 2003; Sutton et al., 1999a; Yokoi & Diedrichsen, 2019), though there is also evidence for an abstraction over sequences such that the relations between sequence elements are represented independently of the elements themselves (Kornysheva et al., 2019; Shima et al., 2007). I unified these descriptions of hierarchical organisation to propose a normative theoretical framework which I could use to investigate whether and how such an organisation would influence decisions between candidate actions. The central interest here was in conflict between hierarchical levels. How do people decide between courses of action that lead to low-level rewards and those that lead to high-level rewards, and how do people evaluate courses of action that accept losses at low levels towards pursuing high-level outcomes? To answer these questions, I started by searching for evidence for the normative theoretical framework proposed to verify that this was a useful lens for investigating behaviour. Having done so, I then aimed to investigate how actions at different hierarchical levels are evaluated by asking two questions. First, I asked whether and how perceived capability to carry out an action influenced willingness to commit to that action in a way that was influenced by its hierarchical level. Second, I asked whether people hold any subjective bias to value higher- or lower-level outcomes more highly due to their hierarchical level alone. By asking these questions, I aimed to produce a complete view on the way in which the hierarchical level of a course of action changes how people estimate its value.

In Chapter 2, I developed a novel method of testing for latent hierarchical structure from low-level action alone and found that a normative framework derived by combining insights from sequential motor control with hierarchical RL provided the best fit to human behaviour. The task centred on testing for an ability to generalise learned structure to produce completely novel sequences of action in a non-Markovian decision space. I demonstrated that this ability to produce novel sequences of action without practice was captured only by a model which included hierarchically organised action, abstract representations of the relations between sequence elements, abstraction of learning about these abstract representations over distinct states, and a preference to explore at high hierarchical levels. Further, the full span of behaviour observed over the entire task was best captured by a transition from a flat (i.e., non-hierarchical) system of behavioural control to this more sophisticated hierarchical model. These findings therefore support the theoretical framework introduced by combining the well formalised hierarchical structure described by hierarchical RL (Botvinick, Niv, et al., 2009; Sutton et al., 1999b) with insights from the study of sequential motor control (Kornysheva et al., 2019; Shima et al., 2007), and they evidence an impressive ability to generalise learned structure to produce novel behavioural routines in novel settings. These findings also highlight an asymmetry in exploration strategies over hierarchical levels that is relevant for my exploration of how hierarchical level influences choice between actions: exploring at high-levels may be more likely to lead to useful discovery, where exploring at low-levels may be redundant.

In Chapter 3, I extended self-efficacy theory (Bandura, 1974, 1977, 1984), which has classically concerned itself with flat, isolated behaviours, to hierarchically organised action. Over two experiments, I demonstrated that the length of a course of action (as a proxy for hierarchical level) influenced choice by distorting self-efficacy judgments. There were two components to this association between length and self-efficacy. First, as the length of a course of action increases, it prescribes more low-level actions, and this increase in the number of low-level actions that must be performed decreases the likelihood of successfully performing all actions involved without a single failure. I found that people are sensitive to this decrease in likelihood of success with increased length. Second, as the length of a course of action increases, it requires that more individual low-level decisions adhere to its prescriptions. This restriction rules out the possibility of acting on changes in the environment should they conflict with the original course of action. This limit placed on an increasing number of future decisions was aversive, and this aversion was sufficient to prompt a deviation in behaviour away from the optimal solution to the task used here – participants were less likely to commit to lengthier courses of action in non-stationary environments than suggested by a normative maximisation of value. In sum, I found in this chapter that the hierarchical level of an action influences perceptions of self-efficacy in ways that dissuade commitment to it.

In Chapter 4, I investigated whether people are biased in how they estimate the value of outcomes that occupy high vs low hierarchical levels. The rationale for such a bias is that temporal discounting (Odum, 2011), which discounts the value of outcomes that are further away in time, will bias against the pursuit of high-level outcomes given that higher-level outcomes are necessarily further away in time than lower-level outcomes. Therefore, without any counteracting force to inflate the value of higher-level outcomes, they may become infeasible to pursue with any consistency. I used a novel delivery task where participants would choose between two policies of action that were of equal objective value but earned an uneven distribution of rewards over hierarchical levels – one earned primarily low-level reward, and the other high-level reward. No evidence was found to support the presence of any general bias to pursue either high- or low-level outcomes. However, by including minimal social cues to attend to specific outcomes, such biases could be manufactured. It seems therefore that there is no force to counteract the effect of temporal discounting at high hierarchical levels, but that social context can go some way to boosting the subjective value of specific outcomes at specific levels. It remains to be seen whether this is specifically an interaction between social context and *hierarchy*, but these findings open the door to research in this area.

In summary, the theoretical framework presented in the introduction of this thesis has proven to provide accurate accounts of human action and has proven useful in guiding investigations of the influence of hierarchical organisation on choice between actions. I have presented novel behavioural methods for empirical research into latent hierarchical structure, and I have used these methods to explore the intricacies of the relationship between how we organise our behaviour and how we decide between them. I have found that we prefer to explore at high hierarchical levels, that we are biased against high-level courses of action because they are difficult to perform and they require that we tie the hands of our future selves, and that while we are not generally biased in how we estimate the value of outcomes at particular hierarchical levels, our social environment can prompt a subjective bias in what we deem most valuable. These findings paint a more complete picture of how we decide how to behave given how we organise our own behaviour.

5.2 Theoretical Implications

5.2.1 Implications for SMC

Previous research into sequential motor control has focussed on either a progressive process of chunking actions and sequencing chunks (and so on) (Lashley, 1951; Yokoi & Diedrichsen, 2019), or on abstraction over sequences to extract the relational structure between sequence elements (Kornysheva et al., 2019; Shima et al., 2007). I here presented evidence for a unification of these two distinct processes into a single system of behavioural control. These two processes together provide a powerful conjunctive coding of action. Such a code was put forward by Kornysheva et al. (2019) to explain the effector-independent and abstract representation of ordinal position they observed in MEG recordings of the human brain during a sequence production task. Abstract representations of the relations between sequence elements (such as their order, e.g., (1st, 2nd, 3rd)) can be combined with representations of specific actions (e.g., turn, push, and pull) to establish specific sequences of those actions that adhere to the represented structure (e.g., (push, turn, pull)). My findings suggest that this same process can be applied to higher-level routines which sequence together not primitive actions, but lower-level sequences of action. This combination of sequencing to produce progressively higher-level representations of behavioural routines and abstraction to abstract away the relations between sequence elements independent of the content of that sequence provides a powerful, flexible, and efficient code for action. This organisation confirms Lashley's (1951) claim that there must be a hierarchical structure and syntax for all movement just as there is a hierarchical structure and syntax for language. It was unclear from these previous findings (Kornysheva et al., 2019; Yokoi & Diedrichsen, 2019) whether and how sequencing of progressively higher-level representations of action and abstraction over sequences of individual actions were integrated in the human brain. I resolve this question here by presenting evidence for a more complete view of how human action is organised that integrates both processes under a single framework.

5.2.2 Implications for Reinforcement Learning

Implications of the research presented here for RL can be divided into those that bear on computational RL as the study of ever-more capable RL algorithms, and those that bear on the intersection between RL and neuroscience.

5.2.2.1 Computational Reinforcement Learning

To solve the scaling problem, which describes the difficulty flat RL methods have with navigating large state an action spaces, computational RL turned to hierarchy. The most popular implementations of hierarchical RL (see the options framework, Sutton et al., 1999a) make use of temporal abstraction to group together sequences of related actions to form progressively higher-level routines of action. This sequencing of lower-level parts to form higher-level behaviours captures only one of the two architectural processes I identified as necessary for immediate acquisition of novel behaviours. Abstraction over sequences, so that the relations between sequence elements are represented independently of the motor content, was also necessary. The inclusion of this process when paired with abstraction of learning over distinct states provided a powerful scheme for learning what to do quickly in novel environments by making use of prior structural knowledge. Computational RL might usefully consider how this form of abstraction might be best implemented and integrated with state abstraction, which is already well established (Abel, 2019; Andre & Russell, 2002; Barto & Mahadevan, 2003). Doing so could lead to advances in the adaptability of hierarchical RL methods to changes in their environment.

It remains to be seen quite how hierarchies of behaviour emerge from experience. The issue of discovering useful high-level routines of behaviour is captured in computational RL by the option discovery problem (Machado et al., 2017; Stolle & Precup, 2002; Sutton et al., 1999b). Many algorithms have been proposed for discovering useful options, and the findings presented in chapter 2 support those that include an intuitive transition from a flat system of behavioural control dependent on memory to a hierarchical system. To establish a hierarchical organisation of the sort put forward in this thesis, an agent must also discover which relations between sequence elements are worth representing, which hierarchical levels are worth exploring, and which states are similar enough such that abstracting learning between them is useful. For adaptive responses to changes in the environment, I found that all these components were necessary, and so to take computational advantage of the immediate acquisition of novel behaviours observed here, computational RL must account for the discovery of the information required by these components.

A final implication for computational RL involves the focus of this thesis on the influence exerted by the way a system organises its actions on the way it decides between them. Several such influences have been discussed here, and the question for computational RL is whether it would be advantageous to consider implementing these biases in normative models of action. Commitment to an ongoing high-level policy of action may result in losses when faced with unexpected and highly valuable rewards that require off-policy actions to attain them. Therefore, in uncertain and variable environments, it may be optimal to apply a penalty to high-level (and therefore lengthy) courses of action (as observed in human behaviour in chapter 3). Similarly, it may be optimal to apply a discounting factor to counteract the effect of temporal discounting applied at very high hierarchical levels (as discussed in chapter 4) if successful execution of the high-level policies involved is very likely. Whether or not such biases should be included is an open question, and it is one that is heavily influenced by how we define *optimal*. However, these hierarchical biases in estimations of value may be important for maximising expected sums of reward over all levels of the behavioural hierarchy, rather than being biased to maximise low-level reward alone.

5.2.2.2 RL and the Brain

It is well established that the human brain makes use of RL-like systems to decide how to act (Niv, 2009). The observation that the scaling problem must therefore pertain in neuroscience as it does in computational RL led to a search for evidence supporting the use of hierarchical RL-like systems in the brain (Botvinick, Niv, et al., 2009). This search was successful, with evidence being found supporting hierarchically organised prediction errors (Diuk, Tsai, et al., 2013; Ribas-Fernandes et al., 2011). This thesis provides behavioural evidence in further support of hierarchical RL-like systems being used to control human action. I also extend the typical hierarchical RL frameworks used to investigate human behaviour by integrating findings from motor control research. This provided more accurate descriptions of human behaviour, and this success argues for further development and investigation of the normative framework put forward. Hierarchy and abstraction are centrally important for how people organise behaviour and how people decide how to act, and hierarchical RL provides a formal lens through which this can be better understood.

Integrating hierarchical RL with the study of sequential motor control brings to light a dual process of establishing hierarchies by breaking down goals and building up sequences. In Chapter 1, I found support for a hierarchical organisation of human action that brought together insights from hierarchical RL, which describes a process of breaking down goals to progressively lower-level sub-goals, and the study of sequential motor control, which describes a process of building up sequences from oft-repeated chunks of action. This may be a relevant insight for how the human brain solves the option-discovery problem and how it learns useful high-level routines of action. As explained by sequential motor control, the human brain might learn to chunk together individual actions that are frequently repeated, and to sequence together chunks in a similar fashion, building up progressively higher-level representations of sequential action. This might also explain how relational representations are discovered – frequent alternation between actions within multiple distinct sequences may well be recognised and alternation as a relational structure thus extracted. Hierarchical RL describes a second process of breaking down goals into progressively lower-level sub-goals with progressively lower-level routines of action aiming to fulfil them. The evidence presented here in favour of a model of behavioural control derived by integrating SMC with HRL lends credibility to the idea that these dual processes may work in collaboration to discover and establish hierarchies of action. Indeed, such a scheme would align well with the long-held idea that there is a movement from high-level and abstract intention to low-level specific action along the rostro-caudal axis of the frontal cortex in the human brain (Badre & D’Esposito, 2009; Koechlin et al., 2003; Koechlin & Summerfield, 2007).

5.2.3 Implications for the Study of Decision Making

Is the way in which the human brain organises behaviour important for how it decides how to behave? Answering this question was a central aim of this thesis, and the empirical evidence presented converges on the idea that yes, the way in which candidate actions are organised can influence decisions between them: in Chapter 2, I demonstrated that people prefer to explore actions at specific hierarchical levels; in Chapter 3, I demonstrated that the hierarchical level of a given course of action influences people's willingness to commit to it; and in Chapter 4, I demonstrated that social context could bias decisions between actions in favour of outcomes at specific hierarchical levels. The hierarchical organisation used by the human brain clearly affects how the human brain decides what to do, and I will now discuss how research into the brain's decision-making processes can and should take this into account.

The trade-off between exploration and exploitation is central to all decisions of how to act (Berger-Tal, Nathan, Meron, & Saltz, 2014; Kayser, Mitchell, Weinstein, & Frank, 2015; Macready & Wolpert, 1998; Mehlhorn et al., 2015). Is it better to explore new courses of action in search of a new optimum, or is it better to exploit learned information by opting for a course of action known to be rewarding? Most solutions to this trade-off centre on establishing at what point during learning one should transition from a primarily explorative strategy to a primarily exploitative one. For a simple example, in RL one can set a small probability of ignoring all learned information and selecting randomly between all available actions, and this small probability can decay with time to indicate a lesser need for exploration with more experience (this is the epsilon-greedy algorithm). The hierarchical perspective taken in this thesis has highlighted a new feature of this trade-off, which is a level-dependency in how it is managed. To accurately describe human behaviour, I found that a preference for high-level exploration was necessary. This could also be a more general level-appropriate mode of exploration, where the brain chooses to explore at a level that somehow matches the environment, but the central insight is the same – when deciding whether to explore or exploit, not all actions are equal. This is most intuitive for very low-level behaviours; there is no use in exploring new methods of reaching for a handle, such as extending the arm behind one's back.

Therefore, the hierarchical organisation used by the human brain to structure behaviour can go some way to simplifying its decision between exploring new courses of action and exploiting those it knows to be rewarding.

Self-efficacy theory (Bandura, 1977) has a long history and a strong body of research supporting the interactions it describes between perceived capability and choice between actions (DiClemente et al., 1995; Maddux & Stanley, 1986; Weinberg, Gould, & Jackson, 2019), though this has not been without debate (Eastman & Marzillier, 1984; Kirsch, 1995; D. M. Williams, 2010). Throughout its history, however, self-efficacy has described actions as isolated, flat behaviours that can all be equivalently compared with respect to one's own perceived capability to carry them out. However, here I demonstrated that this is not the case: higher-level actions are more difficult to perform because they describe lengthier sequences of action which introduces more opportunity for failure. This insight was captured by coping self-efficacy (Chesney et al., 2006) which was an extension of self-efficacy to probe individuals for how well they expected they would be able to maintain adherence to a given policy of sustained action. Although this is a hierarchical idea, hierarchy was only ever implicit in the underlying theory. The results presented here suggest that a wider revision of self-efficacy theory and its applications is necessary to ensure that wherever measures of self-efficacy are taken, they are done so in full knowledge of the relative hierarchical levels of the actions under consideration.

The length of higher-level courses of action had a second effect on behaviour – the requirement that future decisions adhere to a commitment to a high-level course of action, and the restrictions placed on future choices to fulfil this requirement, was aversive. The rationale put forward here to explain this effect is that lengthier restrictions placed by a given high-level policy on low-level choices make it more likely that unexpected and desirable off-policy options are encountered, and in anticipation of this desire to deviate from the high-level policy at some point in the future, people become less likely to commit to the policy in the first place. By contrast with the previous discussion on self-efficacy which captures the mechanical difficulty of executing a sequence of actions, this effect captures the cognitive difficulty of consistent self-restraint. Indeed, cognitive effort has been implicated in the maintenance of cognitive control during goal pursuit (Dayan, 2012; Kurzban, Duckworth, Kable, & Myers, 2013; Shenhav, Botvinick, & Cohen, 2013; Westbrook, Kester, & Braver, 2013), and cognitive effort discounting (that is, a reduction in the subjective value of an outcome based on the cognitive costs borne to attain it) has been observed in human decision making (Botvinick, Huffstetler, & McGuire, 2009; Magno, Foxe, Molholm, Robertson, & Garavan, 2006; Shenhav et al., 2013). These findings and the general framework of cognitive control align with the results presented in this thesis. The prospect of rejecting all future unexpected and desirable off-policy options may represent too large a cognitive cost and thus the value of the high-level policy itself may be discounted by the cognitive effort required to maintain it. Indeed, cognitive control and hierarchy go hand-in-hand – a hierarchical organisation of behaviour requires strict control to adhere to (or else change) all levels of the hierarchy that bear on low-level decisions. The restrictions placed by progressively higher-level policies on choice and the effect these restrictions have on behaviour therefore highlight the need to consider hierarchy when investigating cognitive control (and vice versa).

Temporal discounting reduces the subjective value of outcomes that are further away in time (Green & Myerson, 2004; Green, Myerson, & Macaux, 2005; Green, Myerson, & McFadden, 1997; Odum, 2011), and given that higher-level outcomes are necessarily further away in time, temporal discounting establishes a bias to prefer to pursue lower-level outcomes. In this thesis, I investigated whether any counteracting force was used by the human brain to limit the strength of this bias. I found no evidence in support of any general boost in subjective value for higher-level outcomes (which would counteract the effect of temporal discounting), which indicates that temporal discounting may indeed bias the human brain to prefer lower-level outcomes. However, it may be the case that while there is no global boost in value for higher-level outcomes, the rate of temporal discounting applied to a given outcome depends upon its hierarchical level and that the timescales used in the experiments presented here were too small to detect any meaningful difference in temporal discounting rates. Whether or not there is any mechanism used to balance the asymmetrical effects of temporal discounting over the behavioural hierarchy, it is clear that temporal discounting and hierarchical organisation work together to bias how we decide between candidate actions. If there is conflict between high and low hierarchical levels, temporal discounting will exert a bias in favour of the low level. Understanding whether this is indeed how hierarchy and temporal discounting interact and whether any balancing of the interactions involved is performed is important for a complete view of the effect time has on our decisions.

Social decision-making is a complex field of research that describes many interacting processes (for a review, see Rilling & Sanfey, 2011). In this thesis, I presented evidence for an effect of social context on computations of value, where social cues to attend to specific outcomes biased subjective value in their favour. This aligns with a large body of research describing social biases to value some outcomes over others (see Díaz-Gutiérrez, Alguacil, & Ruz, 2017). However, the hierarchical approach taken in this thesis offers a novel perspective on these biases. If we consider the full behavioural hierarchy followed by any individual person, there are limits on the highest-level goals they may feasibly work towards alone. As a result, a sensible extension of this hierarchy is to include, at levels above one's own highest-level behaviours, an abstraction over the behaviours of other people. Hierarchies could then be assembled over groups of people towards a common goal, which resolves the issue of some high-level goals being unattainable for any one person. Note that this social hierarchy is distinct from social hierarchies as typically studied (e.g., Santamaría-García, Pannunzi, Ayneto, Deco, & Sebastián-Gallés, 2014); this is not a hierarchy of chains of command and superior/inferior individuals, but more simply a hierarchy of behaviours abstracted over multiple people. This mode of social hierarchical behaviour can explain the results presented in this thesis; if I receive a cue that other people are willing to work towards a common goal, then this opens the door for me to collaborate with them. However, it remains to be seen whether the human brain does organise social action in this way, but this is one framework for considering how the hierarchically organised behaviour of an individual could be scaled up to meet the demands of behaving within a group.

In sum, there are many diverse influences of hierarchical organisation on decisions between actions. This is with good reason – a sensitivity to the way in which candidate actions will be implemented and carried out is very likely to be useful in deciding whether to commit to them. Therefore, in answer to the question posed at the outset of this thesis (does the hierarchical organisation of human action influence how people decide between candidate actions?), we may answer: yes, the hierarchical organisation of human behaviour does influence decisions between candidate actions, and the ways in which conflict between hierarchical levels are resolved are many and they are diverse. The degree to which a given hierarchical level is promising for exploration, the length of a given course of action (which correlates with its hierarchical level), the number of decisions over which choice is restricted, the distance of an outcome in time (which also correlates with its hierarchical level), and the social context of an outcome all contribute to both how the value of an outcome and its associated policy of action are computed and the resolution of any conflict between different outcomes.

5.3 Practical Implications

5.3.1 Experimental Implications

If correct, the claim made here that all human action is hierarchically organised has an important implication for experimental design. When measuring behaviour, we must acknowledge that no individual action is ever truly isolated. Our best attempts at isolating behaviour will fail given that all actions performed within an experiment will be made in adherence to a higher-level policy that controls participating in the experiment in the first place. We as neuroscientists must therefore be mindful of the hierarchy of behaviour that sit above the behaviours we intend to measure, and we should consider how these high-level courses of action might influence the lower-level behaviours of interest. In practice, this means both accounting for high-level policies that may be held external to the experiment and which might influence the behaviours we intend to measure, and exploring how the behaviours required within an experiment are organised and what this means for how we take our measurements. Failure to treat hierarchy appropriately here could well lead to poor behavioural measures, misinterpreted results, and flawed frameworks of investigation.

5.3.2 Real-World Applications

The findings presented in this thesis have clear application to real-world intervention design. The hierarchical focus taken here translates particularly well into a description of how best to design interventions that aim to aid in the long-term maintenance of high-level policies. Existing frameworks for behaviour change (such as COM-B and the behaviour change wheel, Michie, van Stralen, & West, 2011) detail how various aspects of cognition (e.g., capability, motivation, emotion etc.) can be integrated to characterise behaviours and interventions and highlight useful areas for intervention. While such frameworks have proven successful (Michie & West, 2013), I here present a formal approach to explaining the ways in which we evaluate high-level behaviour and the ways in which we could therefore intervene to encourage engagement with these high-level behaviours. For example, my finding that the length of a high-level policy of action is a barrier to commitment in two distinct ways leads to the conclusion that simply asking for shorter spans of commitment should increase willingness to commit. For another example, my finding that minimal social cues can boost the subjective value of high-level outcomes leads to the conclusion that establishing a social context to encourage maintenance of a high-level policy will improve adherence to it. These conclusions are not necessarily novel, though the hierarchical perspective taken here does offer a deeper and more complete understanding of the algorithms through which they operate. Further investigations of hierarchically organised behaviour may reveal other important elements of intervention design to aid in many real-world applications.

5.4 Limitations & Future Directions

5.4.1 Methodological Considerations

This thesis focussed exclusively on an analysis of human behaviour to uncover and understand the algorithms that control it. No neuroimaging methods were used, and no claims were made regarding the neural implementations of the hierarchical organisation and effects put forward. This was a deliberate decision – I take the view that careful experimental decomposition of behaviour is best suited for understanding cognitive processes and their component algorithms (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Niv, 2021). Detailed analysis of tasks and the behaviour they elicit has led to much success in identifying the latent cognitive processes that produce behaviour: in low-level perception, scientists correctly inferred from psychophysical experiments that colour vision is implemented by three types of retinal cones and were even able to estimate the cones' wavelength sensitivity (Stiles, 1959); and in higher-level cognition, the role of attentional signals and the information they feed back to low-level perceptual processing areas was correctly predicted by purely behavioural paradigms (Ahissar & Hochstein, 1993; Ahissar & Hochstein, 1997; Hochstein & Ahissar, 2002). By contrast, inferring processes from their neural processors alone is difficult, if at all possible (Krakauer et al., 2017). For example, consider the *Caenorhabditis elegans*, whose behaviour we cannot fully predict despite knowing with precise detail the full circuitry of its 302 neuron nervous system (Bargmann & Marder, 2013). To quote Barlow (2013), “a wing would be a most mystifying structure if one did not know that birds flew”.

The behavioural focus taken here does not imply that application of neuroscientific techniques would not be a fruitful future direction for the research presented. The argument is that a detailed analysis of behaviour is a necessary prerequisite for interpretable investigations of potential implementations of the algorithms uncovered by behaviour (Krakauer et al., 2017; Marr, 1976). More simply, we need to understand the processes that produce behaviour before we can understand how those processes are implemented in the human brain. Consistent with this, a marriage of incisive behavioural experimentation and neuroscientific techniques have led to significant advances in neuroscience. For example, the influential reward prediction error hypothesis of dopamine (Schultz et al., 1997) was derived from a detailed understanding of reward-seeking behaviours paired with precise measurements of midbrain dopaminergic neurons during execution of those behaviours. For other examples, the spatial navigation literature is replete with behavioural paradigms that hint at specific spatial representations that have since been measured in the brain (Banino et al., 2018; Doeller, Barry, & Burgess, 2010; Mittelstaedt & Mittelstaedt, 1980; Moser, Kropff, & Moser, 2008).

This thesis performs a detailed analysis of hierarchically organised behaviour and the interactions between such an organisation and decision-making processes, but I can make no claims as to the neural circuits that underlie the algorithms and effects described. A natural next step, therefore, would be to investigate the underlying circuits. One promising possibility is to extend existing research into the sequencing of action in motor and pre-motor areas to investigate in greater detail the transition from abstract representations of behaviour to precise low-level action along the rostro-caudal axis of the frontal cortex, with the proposed framework in mind. This would mean searching for relational representations of sequences of action, searching for circuits that implement abstraction of learnings about these relational representations across distinct but related states, and searching for circuits that compute the value of candidate high-level actions with all relevant elements uncovered in this thesis taken into account. Doing so would, if successful, refine the framework presented here and lend further support to it as an explanation of how human behaviour is organised.

5.4.2 Conceptual Considerations

5.4.2.1 Social Influences

In Chapter 4, I uncovered that minimal social cues were sufficient to bias choice between outcomes at different hierarchical levels. I have discussed this effect at length, though my treatment of the social component is limited. There are many relevant interactions that feed into social hierarchies and delving in detail into the social component uncovered here is beyond the scope of this thesis. My search in Chapter 4 was for any bias in the subjective evaluation of outcomes that occupy distinct hierarchical levels, and I found such a bias in this effect of social context. However, the effect does require further investigation to be more completely understood, and it is likely that social effects on hierarchical behaviour are a much wider topic that go far beyond the one effect measured here. Future research could aim to understand more completely how interactions between people allow them to establish hierarchies of group behaviour that go beyond the behaviour any one individual could perform alone.

5.4.2.2 A Complete View on Hierarchical Decision-Making

The treatment of hierarchical behaviour presented in this thesis establishes a normative framework for investigations of not only how human behaviour is organised, but also of what this organisation means for how decisions between candidate actions are made. I investigated several ways in which the organisation of behaviour could influence decisions between action, but my investigations are non-exhaustive, and they do not provide a complete view on hierarchical decision-making. Future research could adopt and extend the framework presented here to investigate other influences on action selection and to provide a deeper understanding of how the human brain controls behaviour. I contend that all human action is hierarchically organised. Many established findings in psychology and neuroscience concerning how people decide how to act may benefit from a second pass with a hierarchical lens.

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

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.90.12.5718

Ahissar, Merav, & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*. https://doi.org/10.1038/387401a0

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

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002a). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.162485599

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002b). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13172–13177. https://doi.org/10.1073/pnas.162485599

Badre, D., & D’Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2007.19.12.2082

Badre, D., & D’Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews. Neuroscience*, *10*(9), 659–669. https://doi.org/10.1038/nrn2667

Bandura, A. (1974). Behavior theory and the models of man. *American Psychologist*, *29*(12), 859–869. https://doi.org/10.1037/h0037514

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1984). Recycling misconceptions of perceived self-efficacy. *Cognitive Therapy and Research*. https://doi.org/10.1007/BF01172995

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-Efficacy Beliefs of Adolescents*. https://doi.org/10.1017/CBO9781107415324.004

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*. https://doi.org/10.1038/s41586-018-0102-6

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*. https://doi.org/10.1016/j.neuron.2020.11.024

Bargmann, C. I., & Marder, E. (2013). From the connectome to brain function. *Nature Methods*. https://doi.org/10.1038/nmeth.2451

Barlow, H. B. (2013). Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication*. https://doi.org/10.7551/mitpress/9780262518420.003.0013

Barto, A. G., Konidaris, G., & Vigorito, C. (2014). Behavioral hierarchy: Exploration and representation. *Computational and Robotic Models of the Hierarchical Organization of Behavior*, *9783642398*, 13–46. https://doi.org/10.1007/978-3-642-39875-9_2

Barto, A. G., & Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems: Theory and Applications*. https://doi.org/10.1023/A:1022140919877

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*. https://doi.org/10.1016/j.neuron.2005.05.020

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0095693

Bernstein, N. A. (1967). *The co-ordination and regulation of movements.* Oxford: Pergamon Press.

Binsch, O., Wabeke, T. R., Koot, G., Venrooij, W., & Valk, P. J. L. (2016). *Enhancing Human Resilience : monitoring, sensing, and feedback*. Retrieved from https://www.narcis.nl/publication/RecordID/oai:tudelft.nl:uuid%3A4fea73fb-3c3e-42bf-9b1d-a61e84d8406d

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*. https://doi.org/10.1086/673885

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. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2008.02.009

Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective and Behavioral Neuroscience*. https://doi.org/10.3758/CABN.9.1.16

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

Bourne, P. G. (2008). "Just Say No": Drug Abuse Policy in the Reagan

Administration. In *Ronald Reagan and the 1980s*. https://doi.org/10.1057/9780230616196_4

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*. https://doi.org/10.1037/h0047313

Brown, R., Miller, G. A., Galanter, E., & Pribram, K. H. (2006). Plans and the Structure of Behavior. *Language*. https://doi.org/10.2307/411065

Bruner, J. S. (1973). Organization of early skilled action. *Child Development*. https://doi.org/10.1111/j.1467-8624.1973.tb02105.x

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98. https://doi.org/10.1126/science.1233912

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*.

Buonomano, D. V., & Laje, R. (2010). Population clocks: Motor timing with neural dynamics. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2010.09.002

Burtt, H. E., & Washburn, M. F. (1918). Movement and Mental Imagery. *The Journal of Philosophy, Psychology and Scientific Methods*. https://doi.org/10.2307/2940157

Chesney, M. A., Neilands, T. B., Chambers, D. B., Taylor, J. M., & Folkman, S. (2006). A validity and reliability study of the coping self-efficacy scale. *British Journal of Health Psychology*. https://doi.org/10.1348/135910705X53155

Connor, J. (2007). *Pascal's Wager: The man who played dice with God*.

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-

6613(00)01800-3

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*. https://doi.org/10.1037/0033-295X.113.4.887

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

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E. J., & Rushworth, M. F. S. (2009). Effort-based Cost-benefit valuation and the human brain. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.4515-08.2009

Da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*. https://doi.org/10.1038/nature25457

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*. https://doi.org/10.1016/j.neuron.2011.02.027

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*. https://doi.org/10.1016/j.conb.2012.05.011

Díaz-Gutiérrez, P., Alguacil, S., & Ruz, M. (2017). Bias and control in social decision-making. In *Neuroscience and Social Science: The Missing Link*. https://doi.org/10.1007/978-3-319-68421-5_3

DiClemente, C. C., Fairhurst, S. K., & Piotrowski, N. A. (1995). *Self-Efficacy and Addictive Behaviors*. https://doi.org/10.1007/978-1-4419-6868-5_4

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding,. In *PLoS Computational Biology*.

Diuk, C., Schapiro, A., Córdova, N., Ribas-Fernandes, J., Niv, Y., & Botvinick, M. M. (2013). Divide and Conquer: Hierarchical Reinforcement Learning and Task Decomposition in Humans. In *Computational and Robotic Models of the*

Hierarchical Organization of Behavior (pp. 271–291). https://doi.org/10.1007/978-3-642-39875-9_12

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

Do, A. M., Rupert, A. V., & Wolford, G. (2008). Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/PBR.15.1.96

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*. https://doi.org/10.1038/nature08704

Doyle, J. R. (2013). Survey of time preference, delay discounting models. *Judgment and Decision Making*. https://doi.org/10.2139/ssrn.1685861

Eastman, C., & Marzillier, J. S. (1984). Theoretical and methodological difficulties in Bandura's self-efficacy theory. *Cognitive Therapy and Research*. https://doi.org/10.1007/BF01172994

Ebbinghaus, H. (2013a). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*. https://doi.org/10.5214/ans.0972.7531.200408

Ebbinghaus, H. (2013b). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*. https://doi.org/10.5214/ans.0972.7531.200408

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2007). Evolutionary development of hierarchical learning structures. *IEEE Transactions on Evolutionary Computation*. https://doi.org/10.1109/TEVC.2006.890270

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*. https://doi.org/10.1037/0033-295X.87.6.477

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories.

Theory and Decision. https://doi.org/10.1007/BF00134215

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*. https://doi.org/10.1037/h0055392

Fitts, P. M. (1964). Perceptual-Motor Skill Learning. *Categories of Human Learning*.

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

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Todd, P. M. (1999). Fast and Frugal Heuristics: The Adaptive Toolbox. In *Simple Heuristics That Make Us Smart*.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*. https://doi.org/10.1016/j.neuron.2010.04.016

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Gollwitzer, P. M., Wieber, F., Myers, A. L., & McCrea, S. M. (2010). How to Maximize Implementation Intention Effects. In *Then A Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research*. https://doi.org/10.1093/acprof:oso/9780195377798.003.0008

Grafman, J. (2002). The human prefrontal cortex has evolved to represent components of structured event complexes. In *Handbook of Neuropsychology*.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.130.5.769

Green, L., Myerson, J., & Macaux, E. W. (2005). Temporal discounting when the choice is between two delayed rewards. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.31.5.1121

Green, L., Myerson, J., & McFadden, E. (1997). Rate of temporal discounting decreases with amount of reward. *Memory and Cognition*. https://doi.org/10.3758/BF03211314

Greenfield, P. M., Nelson, K., & Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: A parallel between action and grammar. *Cognitive Psychology*. https://doi.org/10.1016/0010-0285(72)90009-6

Grossberg, S. (1978). A Theory of Human Memory: Self-Organization and Performance of Sensory-Motor Codes, Maps, and Plans. In *Progress in Theoretical Biology*. https://doi.org/10.1016/b978-0-12-543105-7.50013-0

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained Memory: Error Patterns Rule out Chaining Models of Immediate Serial Recall. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*. https://doi.org/10.1080/713755612

Hirschman, E. C., Kahneman, D., Slovic, P., & Tversky, A. (1983). Judgement under Uncertainty: Heuristics and Biases. *Journal of Marketing Research*. https://doi.org/10.2307/3151689

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*. https://doi.org/10.1016/S0896-6273(02)01091-7

Holden, G. (1992). The Relationship of Self-Efficacy Appraisals to Subsequent Health Related Outcomes. *Social Work in Health Care*. https://doi.org/10.1300/j010v16n01_05

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*. https://doi.org/10.1038/1124

Hoshi, E., Shima, K., & Tanji, J. (1998). Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.1998.80.6.3392

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In *Current Research in Natural Language Generation*.

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

Hyde, J., Hankins, M., Deale, A., & Marteau, T. M. (2008). Interventions to increase self-efficacy in the context of addiction behaviours: A systematic literature review. *Journal of Health Psychology*. https://doi.org/10.1177/1359105308090933

Ibsen, H. (1891). *Hedda Gabler*.

Isbell, J., Shelton, C. R., Kearns, M., Singh, S., & Stone, P. (2001). A social reinforcement learning agent. *Proceedings of the International Conference on Autonomous Agents*. https://doi.org/10.1145/375735.376334

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., … Casey, B. J. (2011). Behavioral and neural properties of social reinforcement learning. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

Kacelnik, A. (1997). Normative and descriptive models of decision making: Time discounting and risk sensitivity. *CIBA Foundation Symposia*. https://doi.org/10.1002/9780470515372.ch5

Kahneman, D., & Tversky, A. (2018). Prospect theory: An analysis of decision under risk. In *Experiments in Environmental Economics*. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (2019). Choices, values, and frames. In *Choices, Values, and Frames*. https://doi.org/10.1017/CBO9780511803475.002

Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychological Assessment*. https://doi.org/10.1037/a0025174

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology*. https://doi.org/10.1038/npp.2014.193

Kirsch, I. (1995). Self-efficacy and outcome expectancies: a concluding commentary. *Self-Efficacy, Adaption, and Adjustment: Theory, Research and Application*, 331–345.

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

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2007.04.005

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

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*. https://doi.org/10.1016/j.neuron.2016.12.041

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X12003196

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2023

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

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

Leonard, J. A., & Newman, R. C. (1964). Formation of higher habits [48]. *Nature*. https://doi.org/10.1038/203550b0

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.1992.67.1.145

Loewenstein, G. F. (1988). Frames of Mind in Intertemporal Choice. *Management Science*. https://doi.org/10.1287/mnsc.34.2.200

Machado, M. C., Bellemare, M. G., & Bowling, M. (2017). A laplacian framework for option discovery in reinforcement learning. *34th International Conference on Machine Learning, ICML 2017*.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on Evolutionary Computation*. https://doi.org/10.1109/4235.728210

Maddux, J. E., & Stanley, M. A. (1986). Self-Efficacy Theory in Contemporary Psychology: An Overview. *Journal of Social and Clinical Psychology*. https://doi.org/10.1521/jscp.1986.4.3.249

Magno, E., Foxe, J. J., Molholm, S., Robertson, I. H., & Garavan, H. (2006). The anterior cingulate and error avoidance. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0369-06.2006

Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. https://doi.org/10.1098/rstb.1976.0090

McGovern, E. A. (2002). Autonomous discovery of temporal abstractions from interaction with an environment. *Power*.

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

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*. https://doi.org/10.1186/1748-5908-6-42

Michie, S., & West, R. (2013). Behaviour change theory and evidence: A presentation to Government. *Health Psychology Review*. https://doi.org/10.1080/17437199.2011.649445

Miller, G. A., Galanter, E., & Pribram, K. H. (2004). The integration of plans. In *Plans and the structure of behavior.* https://doi.org/10.1037/10039-007

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

Mittelstaedt, M. L., & Mittelstaedt, H. (1980). Homing by path integration in a mammal. *Naturwissenschaften*. https://doi.org/10.1007/BF00450672

Moritz, S. E., Feltz, D. L., Fahrbach, K. R., & Mack, D. E. (2000). The relation of self-efficacy measures to sport performance: A meta-analytic review. *Research Quarterly for Exercise and Sport*. https://doi.org/10.1080/02701367.2000.10608908

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain’s spatial representation system. *Annual Review of Neuroscience*. https://doi.org/10.1146/annurev.neuro.31.061307.090723

Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2006.18.6.974

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.3001028

Nezlek, J. B., & Forestell, C. A. (2020). Vegetarianism as a social identity. *Current Opinion in Food Science*. https://doi.org/10.1016/j.cofs.2019.12.005

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

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*. https://doi.org/10.1037/bne0000471

Odum, A. L. (2011). DELAY DISCOUNTING: I'M A K, YOU'RE A K. *Journal of the Experimental Analysis of Behavior*. https://doi.org/10.1901/jeab.2011.96-423

Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. *Neural Information Processing Systems (NIPS)*.

Paslakis, G., Richardson, C., Nöhre, M., Brähler, E., Holzapfel, C., Hilbert, A., & de Zwaan, M. (2020). Prevalence and psychopathology of vegetarians and vegans – Results from a representative survey in Germany. *Scientific Reports*. https://doi.org/10.1038/s41598-020-63910-y

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

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

Rhodes, B. J., & Bullock, D. (2002). A scalable model of cerebellar adaptive timing and sequencing: The recurrent slide and latch (RSL) model. *Applied Intelligence*. https://doi.org/10.1023/A:1015736004189

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. A. (2004a). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*.

https://doi.org/10.1016/j.humov.2004.10.008

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

Ribas-Fernandes, J. J. F., 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. https://doi.org/10.1016/j.neuron.2011.05.042

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*. https://doi.org/10.1146/annurev.psych.121208.131647

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.1990.63.3.592

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

Rosenfeld, D. L. (2018). The psychology of vegetarianism: Recent advances and future directions. *Appetite*. https://doi.org/10.1016/j.appet.2018.09.011

Rosenstock, I. M., Strecher, V. J., & Becker, M. H. (1988). Social Learning Theory and the Health Belief Model. *Health Education & Behavior*. https://doi.org/10.1177/109019818801500203

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.07.009

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

Santamaría-García, H., Pannunzi, M., Ayneto, A., Deco, G., & Sebastián-Gallés, N. (2014). “If you are good, i get better”: The role of social hierarchy in perceptual decision-making. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nst133

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

Sayette, M. A. (2004). Self-regulatory failure and addiction. In *Handbook of self-regulation: Research, theory, and applications(1st Edition)*.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*. https://doi.org/10.1037/h0076770

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*. https://doi.org/10.1523/jneurosci.13-03-00900.1993

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

Schwarzer, R., & Renner, B. (2000). Social-cognitive predictors of health behavior: Action self-efficacy and coping self-efficacy. *Health Psychology*. https://doi.org/10.1037/0278-6133.19.5.487

Shallice, T., & Burgess, P. W. (1991). Higher-order cognitive impairments and frontal lobe lesions in man. In *Frontal love function and dysfunction*.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013a). Cortical control of arm movements: A dynamical systems perspective. *Annual Review of Neuroscience*.

https://doi.org/10.1146/annurev-neuro-062111-150509

Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013b). Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience*. https://doi.org/10.1146/annurev-neuro-062111-150509

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

Şimşek, Ö., Wolfe, A. P., & Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*. https://doi.org/10.1145/1102351.1102454

Singh, S., Barto, A. G., & Chentanez, N. (2005). Intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*.

Skinner, B. F. (1935). Two Types of Conditioned Reflex and a Pseudo Type. *The Journal of General Psychology*. https://doi.org/10.1080/00221309.1935.9920088

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. *Neuron*. https://doi.org/10.1016/j.neuron.2019.06.012

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

Spek, V., Lemmens, F., Chatrou, M., Van Kempen, S., Pouwer, F., & Pop, V. (2013). Development of a smoking abstinence self-efficacy questionnaire. *International Journal of Behavioral Medicine*. https://doi.org/10.1007/s12529-012-9229-2

Stajkovic, A. D., & Luthans, F. (1998). Self-Efficacy and Work-Related Performance: A Meta-Analysis. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.124.2.240

Stiles, W. S. (1959). COLOR VISION: THE APPROACH THROUGH INCREMENT-

THRESHOLD SENSITIVITY. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.45.1.100

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. (1999a). 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

Sutton, R. S., Precup, D., & Singh, S. (1999b). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in RL. *Statewide Agricultural Land Use Baseline 2015*.

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.00238.2004

Terrace, H. S. (2005). The simultaneous chain: A new approach to serial learning. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2005.02.003

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases Amos Tversky; Daniel Kahneman. *Science*.

von Neumann, J., & Morgenstern, O. (2007a). Theory of games and economic behavior. In *Theory of Games and Economic Behavior*. https://doi.org/10.2307/2019327 von Neumann, J., & Morgenstern, O. (2007b). Theory of games and economic behavior. In *Theory of Games and Economic Behavior*. https://doi.org/10.2307/2981222

Washburn, M. F., & Thorndike, E. L. (1912). Animal Intelligence: Experimental Studies. *The Journal of Philosophy, Psychology and Scientific Methods*. https://doi.org/10.2307/2013764

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

Watson, J. B. (1920). IS THINKING MERELY ACTION OF LANGUAGE MECHANISMS? (V.). *British Journal of Psychology. General Section*. https://doi.org/10.1111/j.2044-8295.1920.tb00010.x

Wayne Aldridge, J., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: A “natural action” approach to movement sequence. *Journal of Neuroscience*. https://doi.org/10.1523/jneurosci.18-07-02777.1998

Weinberg, R., Gould, D., & Jackson, A. (2019). Expectations and Performance: An Empirical Test of Bandura’s Self-efficacy Theory. *Journal of Sport Psychology*. https://doi.org/10.1123/jsp.1.4.320

Weiss, D. J., & Shanteau, J. (2021). The futility of decision making research. *Studies in History and Philosophy of Science*. https://doi.org/10.1016/j.shpsa.2021.08.018

Westbrook, A., Kester, D., & Braver, T. S. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0068210

Whittington, J. C. R., Muller, T. H., Mark, S., Barry, C., & Behrens, T. E. J. (2018). *Generalisation of structural knowledge in the hippocampal-entorhinal system*. (Nips). Retrieved from http://arxiv.org/abs/1805.09042

Wickelgren, W. A. (1969). Context-Sensitive Coding in Speech Recognition, Articulation and Development. In *Information Processing in The Nervous*

System. https://doi.org/10.1007/978-3-642-87086-6_5

Williams, D. M. (2010). Outcome expectancy and self-efficacy: Theoretical implications of an unresolved contradiction. *Personality and Social Psychology Review*, *14*(4), 417–425. https://doi.org/10.1177/1088868310368802

Williams, S. L. (1995). *Self-Efficacy, Anxiety, and Phobic Disorders*. https://doi.org/10.1007/978-1-4419-6868-5_3

Wise, R. A., Spindler, J., Dewit, H., & Gerber, G. J. (1978). Neuroleptic-induced "anhedonia" in rats: Pimozide blocks reward quality of food. *Science*. https://doi.org/10.1126/science.566469

Wise, R. A., Spindler, J., & Legault, L. (1978). Major attenuation of food reward with performance-sparing doses of pimozide in the rat. *Canadian Journal of Psychology*. https://doi.org/10.1037/h0081678

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience*. https://doi.org/10.1038/nrn1033

Yerkes, R. M., & Morgulis, S. (1909). The method of Pavlov in animal psychology. *Psychological Bulletin*. https://doi.org/10.1037/h0070886

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