The Nature of Representation in Cognitive Control

Dorian Minors

Bachelor of Psychology (Honours), Macquarie University

Author Note

Dorian Minors, Macquarie University. Supervised by Doctor Colin Klein, Department of Philosophy, Macquarie University and Associate Professor Andrew Barron, Department of Biology, Macquarie University.

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Research in the Department of Philosophy, Macquarie University, 2017.

Table of Contents

List of Figures

Abstract

Cognitive control broadly refers to those processes which adaptively coordinate behaviour in service of a goal. To achieve control, the brain must resolve conflicting information and competing cognitive demands, even when doing so runs counter to more dominant, or prepotent impulses. Explaining this property in the context of the brain has long posed a general problem to researchers. Mechanisms of control have been typically posed as intentional processes and are thus subject to anthropomorphism—styled as a brain within the brain. It is difficult to imagine how neural circuits can achieve this. Classical cognitive science has often been criticised for invoking these 'homunculi' to account for control-related processing. Contemporary neuroscientific and computational literature provides an opportunity to resolve these homuncular accounts. Neural network function provides a plausible means of representing information in the brain. Viewed through the lens of network dynamics, certain structural and functional specialisations characterising control-related phenomena can be grounded in neurally plausible properties of the brain. I pay particular attention to how the circuit organisation of the neocortex may contribute to cognitive control mechanisms. I show that such a structure would achieve a high level of cognitive control as an emergent property of network function, without the need to invoke homuncular mechanisms.

Keywords: representation, cognitive control, neural networks, neocortex

Statement of Candidate

I certify that the work in this thesis entitled 'The Nature of Representation in Cognitive Control' has not previously been submitted in any capacity for a degree at Macquarie University, nor has it been submitted as part of requirements for a degree to any university or institution other than Macquarie University.

This thesis is an original piece of research and has been written by me. Any assistance I have received in my research work and the preparation of the thesis has been appropriately acknowledged.

All information sources and literature used have been indicated in the thesis.

Dorian Minors

Neuroscience has achieved a great deal in identifying neural mechanisms that can account for the straightforward transformation of environmental input into behavioural output (Kandel, 2009; Miller, 2000). Yet the behavioural repertoires of many organisms far exceed simple stimulus-response relationships. Humans and other sophisticated animals are capable of organising behaviour to achieve goals that are removed from any proximal circumstance (e.g. Craik, 1967; Hull, 1943; Premack & Premack, 1983; Tolman, 1951). To do this, animals must resolve conflicts between competing cognitive demands, particularly when an adaptive response runs counter to more dominant, or prepotent impulses (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick & Cohen, 2014). This ability to resolve internal conflicts is typically referred to as cognitive control, and it is not yet understood how such a complex process is accomplished by the nervous system (Koechlin, Ody, & Kouneiher, 2003; Miller, 2000; Miller & Cohen, 2001). Questions of how, and to what degree, information about the world might be preserved, or 'represented' in the brain to support this complexity have generated a great deal of useful conjecture in both philosophy and cognitive science. Yet, only recently have we gained the means to model one such interpretation in the context of the nervous system (Clark, 1995). A connectionist perspective provides a neurally plausible account of representation that offers a novel means to investigate cognitive control (Botvinick et al., 2001; Cohen, Dunbar, & McClelland, 1990; Collins & Frank, 2013). Exploring how far such a perspective can take us will set the stage for a new and exciting foray into neural origins of complex behaviour. My thesis asks the question, how can we understand the phenomena of cognitive control in a manner that is compatible with a neurally plausible perspective on representation?

Methodology

Before we begin, I open with a broad outline of my methodology. This thesis is, at its core, a literature review in light of a certain hypothesis. In this respect, my method is broadly

based on methodological naturalism. The phenomena of interest are part of the natural world, and as such any investigation which aims to comment on their nature should be continuous with the sciences. Thus, I will be looking for how these phenomena and their associated terms or concepts are employed by working scientists, as opposed to engaging in something more like a traditional conceptual analysis (e.g. Jackson, 1998). Here, "[t]he point…is not to develop conceptual truths about minds, but rather to deal with philosophical issues through close attention to developments in…cognitive science" (Gabbay, Woods, & Thagard, 2006, p. x; see also Hartner, 2013).

While the more narrow aspects of my approach, along with the premise and background to this research project are more comprehensively explored in chapter 1, the broad thrust is as follows. To explore my research question, I analyse contemporary work on cognitive control with particular reference to cognitive philosophical, psychological, neurobiological, neuroethological, and computational literature. I take my cues from Barbara Webb (2006), Andy Clark (1995), Matthew Botvinick and Jonathan Cohen (2014), James McClelland (1988), Earl Miller (2000), and Michael Graziano (2016) who have each contributed foundational works in the domain and have been used to structure the premise of my arguments. A comprehensive literature review ensued, consolidating and evaluating perspectives from these myriad sources to explore my hypotheses in detail. The content of the study is presented in the form of a chapterised thesis.

Prospectus

Chapter 1. I will suggest that representation should be distinguished as a special function of the nervous system in cases where neural activity somehow 'stands in' for aspects of the environment to produce behaviour in the absence of an eliciting stimulus. Viewed this way, I argue that representation can be considered as a feature of neural network function,

with implications for the kinds of information that may be preserved in this context and thus, the mechanisms for control.

Chapter 2. I will go on to propose that certain structural and functional specialisations characterising control-related phenomena are best understood in terms of this neurally plausible form of representation. Specifically, an account of neural conflict resolution can be largely explained in terms of the adaptive tuning of neural networks to attain, or maintain desirable states of the world and avoid that which might obstruct those states.

Chapter 3. However, such a view cannot explain more complex goal-directed behaviour. Sophisticated animals appear to develop intricate domain-general representations that can adapt the parameters of information processing to achieve control in more complex tasks. I will argue that traditional accounts of this phenomenon are insufficient, and that a more productive approach would be to consider these representations as an epiphenomenal product of the interaction between environmental structure and the neural systems which process it. I suggest that the neocortex possesses the architectural features necessary to plausibly accomplish this as a property of network function. Moreover, I demonstrate that this network structure would achieve a high level of cognitive control organically, without the need to appeal to additional mechanisms. I will conclude by highlighting the features of control phenomena that continue to defy explanation, and outline future avenues of empirical pursuit in relation to the internal origination of complex behaviour.

Chapter 1. Representation in the Brain

1.1 Representation in Cognitive Science

Representation at its most philosophically uncomplicated involves things that 'stand in' for other things, as a painting might stand for its subject, or a lawyer stands for her client (Haugeland, 1991; Kalhat, 2015). The notion of representation was adopted by cognitive science to account for the observation that not all animal behaviour can be explained by appealing to the reinforcement history of some eliciting stimulus (see Craik, 1967; Tolman, 1951; although c.f. Hull, 1943; Skinner, 1938). For circumstances in which a stimulus is not present to stimulate a response, brains must 're-present' the stimulus internally to elicit that same response; neural activity must somehow 'stand in' for aspects of the environment to produce behaviour when the evoking stimulus is not available (Webb, 2006). Cognitive science has typically concerned itself with symbolic mental objects which explicitly represent information about the world and can be stored, retrieved and acted on to make decisions (e.g. Chomsky, 2000; Fodor, 1975, 1983; Marr, 1982; Newell & Simon, 1976).

However, this view has proven problematic, as our models of neural architecture do not provide an intuitive method of carrying information about the world forward in an enduring form (Gallistel, 2006). This issue has been exacerbated by the widespread and uncritical conflation of the terms 'encoding' and 'transformation' with 'representation' by neuroscientists (Webb, 2006). Associative learning is characterised by changes in neural structure which elicit behaviours in response to stimuli that correspond to predictable outcomes (Bienenstock, Cooper, & Munro, 1982; Hebb, 1949; Kandel, 2009). In such cases, while it is generally possible to view the neural activity that correlates with a stimulus as a code which preserves information in some way, the code itself is not necessarily representing. The nervous system is not literally trying to reconstruct the stimulus in the form of a neural code, nor is the encoded information 'decoded' later on, as one might

decrypt a message from an encrypted source (Webb, 2006). Rather, this information forms part of a cascade of internal processes, the role of which is to *transform* the stimulus into a behavioural response. These transformations actually retain less information about the stimulus as the system propagates toward response specification (Webb, 2012). To call these 'representations' merely because they correspond to some environmental event is to confuse the process which links a stimulus to behaviour with something the animal can use as a 'stand-in' for the world (Webb, 2006, 2012). In many cases describing this process does nothing to explain circumstances in which the nervous system does represent—that is, when it produces internal states that can be used in some equivalent way as external stimuli in order to act when the stimuli aren't present.

This, appraised in combination with models in computational literature that mimic properties of complex behaviour without requiring representations, has led some to suggest that we should try to understand the brain without appeal to representations (e.g. Chemero, 2000; Dennett, 1987; Hutto & Myin, 2013). However, I contend that there is no need for an absolutist anti-representational stance, so long as we are clear about what kinds of representations are possible in the context of the nervous system (Clark & Toribio, 1994; Gallistel, 2008).

1.2 Brains Can Represent

A connectionist account provides a viable mechanism for a certain form of neural representation (Clark, 1995). *Neural network modelling* refers to a connectionist approach to computation that attempts to mimic the properties of neurons organised as a network (Bechtel & Abrahamsen, 1991; Churchland, 1989; Feldman, 1981; LeCun, Bengio, & Hinton, 2015; McClelland, 1988; Rumelhart, McClelland, & PDP Research Group, 1986). Inputs into the system are encoded as an activation pattern of 'input units', or artificial neurons. This

initial signal is transformed by associative layers of 'hidden units' into an 'output unit' activation pattern that specifies the response (Fig. 1). The connections between units are given a weight value, affecting the likelihood that the connected unit will pass the signal on. If the response is inappropriate to the input, then the weights are algorithmically adjusted so that the activation pattern of output units elicited by the activation pattern of input units is more suitable, mimicking neural associative learning mechanisms (Bechtel & Abrahamsen, 1991; Hinton, McClelland, & Rumelhart, 1986). In this way, the weighted connections between units determine which inputs prompt which outputs. Here lies the representation, but in a non-intuitive form (Clark, 1995, 1998). The connections contain information that can 'stand-in' for the input, but that information is simply how the signal should be passed

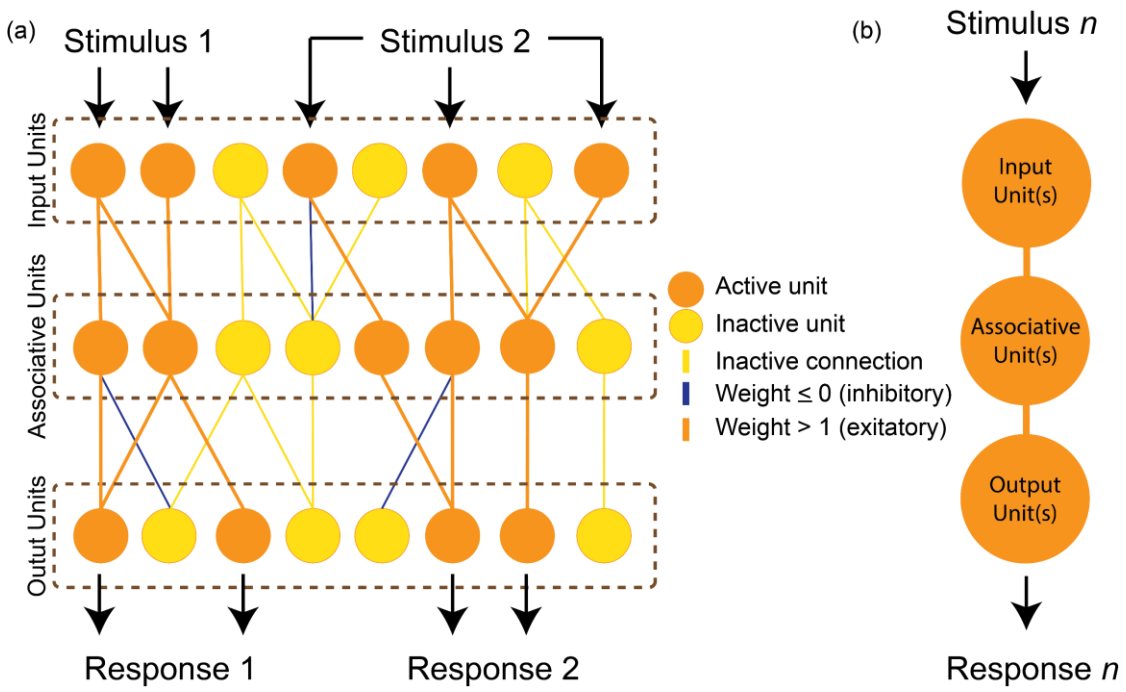

Figure 1. Diagram of a generic neural network architecture. (a) Stimuli are encoded as an activation pattern of 'input units', or artificial neurons. This initial signal is transformed by associative layers of 'hidden units' into an 'output unit' activation pattern that specifies the response. The weight value of connections between units determines whether the signal will be passed on or not. The weight values are determined by an algorithm which attempts to match inputs to a prespecified output. In a biological organism, these weights would be adjusted by Hebbian synaptic plasticity. (b) Commonly, these input-output pathways are illustrated in this way, with a single unit used to represent the activation patterns in each layer. Alternatively, the pathway may be displayed without an associative unit.

on. The connection weights describe what outputs are possible in the context of specific input activation patterns (Hinton et al., 1986).

We must return to describing organisms for the import of this to become clear, a feat made possible by the similarity between neural network models and the functionality of the brain (London & Häusser, 2005; Marder & Thirumalai, 2002). If we consider that the activation patterns of input neurons in animals would be triggered by stimuli in the environment, and consequent responses are triggered by the activation patterns of connected output neurons, then these neural networks 'stand in', not for the world, but for the responses possible in the environmental context associated with an input neuron activation pattern (Cisek, 2001; Clark, 1995; Engel, Maye, Kurthen, & König, 2013; Millikan, 1995). This means that to encode a stimulus is to at once compute the appropriate response(s). So long as we can elicit the input neuron activation pattern, even if the stimulus normally responsible for that pattern is absent, the animal can use the information contained within connections between neurons. The inner state of the animal can act as a proxy for the world. Indeed, for the purpose of illustration, they could be viewed as explicit representations of Gibson's (1979) 'affordances' in that the inner states activated stipulate the actions available or 'afforded' to an organism in certain environmental contexts (Cisek, 2001; Clark, 1995).

That such things are truly representations remains an open question in philosophical terms; these are different to traditional accounts in many ways. Yet, these '*neural network representations'* inarguably satisfy the premise upon which the term representation was introduced to cognitive science: they present a plausible opportunity for neural activity to 'stand in' for an eliciting stimulus regardless of its physical presence. More significantly, this account of representation allows us to explore the thorny topic of cognitive control.

1.3 Cognitive Control and Associated Phenomena

Very broadly, cognitive control refers to those processes which flexibly coordinate internal processes and behaviours in service of a goal (Cooper, 2010; Koechlin et al., 2003; Miller, 2000; Posner & Snyder, 1975; Shiffrin & Schneider, 1977).[1] While the circumstances purported to require control are somewhat amorphous in the literature, the most demonstrable are those which require the animal to resolve conflicts between competing cognitive demands. This is particularly evident where more reflexive or dominant propensities run counter to the goal. This phenomenon is clearly illustrated by a typical variation of the Stroop task (Fig 2.; also see Stroop, 1935).

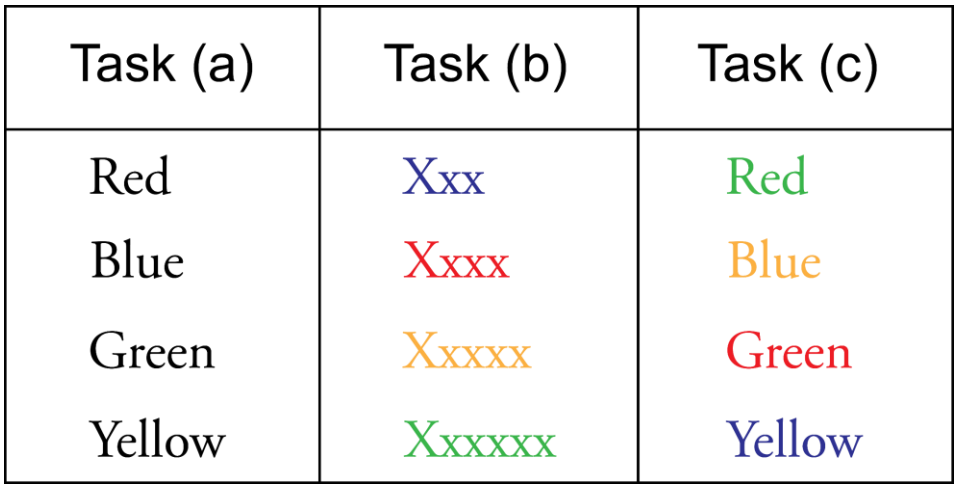

Figure 2. A typical variation on the Stroop task (Stroop, 1935). In Task (a), participants are required to read the words. In Tasks (b) and (c), participants are required to name the colour of the ink. Task (c) is typically considered to be more difficult than tasks (a) or (b) and appears to require effortful control to execute.

In this task, participants are asked to name the colour of the ink a word is printed in. When the word describes a non-corresponding colour, for example the word 'blue' printed in red ink, the task is generally more difficult than simply reading the words, evidenced by increased error rates and response times (e.g. Posner & Snyder, 1975). The presence of some control-related process is relatively easy to identify in this context, in the form of an

[1] As distinct from stimulus-control, in which habitual responding determines behaviour. Though it should be noted that the distinction between habitual or reflexive behaviours, and those which are truly purposeful is not always clear (see for example Dickinson, 1985; Frijda, 2016).

experience of effortful attention in executing the task. One must sustain focus on the colours to avoid reading the words. Here, the cognitive scientist would say that the dominant, prepotent response (word-reading) is 'controlled' such that the less dominant response (colour-naming) can be executed.

Yet, despite a similarly conspicuous presence in many complex tasks, the nature and composition of control-related phenomena have proven elusive (Botvinick et al., 2001; Cooper, 2010; Koechlin et al., 2003; Miller, 2000; Miller & Cohen, 2001).

1.3.1 The homuncular problem of cognitive science. This difficulty forms the foundation upon which the construct of the '*executive'* was introduced into cognitive science. The executive comprises a hypothetical set of functions which arbitrate over lower order processes to facilitate goal-directed behaviour (see for example Baddeley & Hitch, 1974; Barkley, 1997; Miyake et al., 2000; Norman & Shallice, 1980; Zelazo, Carter, Reznick, & Frye, 1997). Yet, naming these arbitrators does little to explain them and we are still faced with the problem of mapping these processes to neural function (Botvinick & Cohen, 2014; McClelland et al., 2010). Too often researchers will uncritically ascribe cognitive processes, or their deficits, to the executive without questioning their mechanistic nature. Performance deficits in the Stroop, for example, tend to be attributed to a lapse in a hypothetical supervisory attentional system (Badgaiyan, 2000; Norman & Shallice, 1980; Stuss, Shallice, Alexander, & Picton, 1995). As with any hypothetical construct, this presents the risk of reification; coming to believe that naming these functions somehow explains their performance (Hull, 1943).[2]

[2] Alternatively, it is subject to a doctrine of emergentism; a position that the processes are unexplainable in terms of more basic properties.

One must also consider the corollary: over what are these arbitrators arbitrating? With little clarity around what information is actually preserved in the context of the nervous system, this problem appears insurmountable. Grounded as they are in symbolic notions of mental content, executive functions are typically considered to reflectively arbitrate over detailed internal models of the world. Under these conditions, the executive becomes an ill-defined set of homunculi—unexplained intelligences, or 'little men' in the brain—which are treated as the causal source and solution to many cognitive problems, including that of control (Botvinick & Cohen, 2014; Hazy, Frank, & O'Reilly, 2007; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). The assumption here is that the homuncular mechanism looks at the model of the world and makes the appropriate adjustments. We are left to explain how the homunculus is making its decisions, which is the core of the original problem all over again (Gregory, 1969; Hull, 1943). Empirical predictions made on this basis are likely to lead us astray.

However, deferring explanations of control to the executive has allowed cognitive science to narrow the scope of the empirical problem. The traditional neurocognitive literature presents a number of 'executive' functions thought to facilitate conflict resolution; the ability to maintain sustained attention (Pennington & Ozonoff, 1996; Smith & Jonides, 1999); to plan steps toward a goal (Duncan, 1986; Passingham, 1993; Shallice, 1982; Smith & Jonides, 1999); to subsequently initiate goal directed behaviour (Lezak, 2004); to hold, integrate, and manipulate information in the mind over time to support appropriate response selection (i.e. 'working memory', Fuster, 1988; Fuster & Alexander, 1971; Goldman-Rakic, 1987, 1996; Goldman-Rakic, 1987); and inhibit inappropriate responses (Luna, Padmanabhan, & O'Hearn, 2010; Smith & Jonides, 1999); to monitor performance (Petrides & Milner, 1982); to learn task 'rules' to support performance (Fantino, 2003; Miller, Nieder,

Freedman, & Wallis, 2003; Seger & Miller, 2010); and finally, the ability to shift between goal states (Ravizza & Carter, 2008).[3]

These functional distinctions have been supported by the discovery of neural systems that apparently correspond to these executive mechanisms, primarily associated with cortical regions of the brain such as the pre-frontal cortex (Luria, 1970; Miller, 2000; Niendam et al., 2012; Shallice, 1988) or their analogues in animals which lack a cortex (e.g. Diekamp, Kalt, & Güntürkün, 2002; Güntürkün, 2005). Yet the components of the executive, in their current form, resist attempts to plausibly map them onto neural architecture (Botvinick & Cohen, 2014; McClelland et al., 2010). So long as cognitive science models these processes at the level of the cognitive agent, there can be no unification of these theories of control with tenets of neuroscience, which tends to model instead at the level of the neural circuit.

There is a question as to whether control is in fact reduceable to some number of component processes. The idea that control should instead be viewed as an emergent product of other cognitive functions (an argument clearly articulated in Cooper, 2010) has seen several treatments and in fact stretches back to the ostensible birth of the domain of study. In his seminal analysis of goal-directed behaviour, Thorndike wondered whether the capacity to control impulses was merely the consequence "of an increase in the number, delicacy, and complexity of associations of the general animal sort" (1911, p. 286). Unfortunately, answering this question has proven impossible in the absence of a plausible account of representation, and the empirical data in this direction is lacking.

As such, I will consider how the hypothetical components of cognitive control outlined above might be supported by neurally plausible mechanisms. In doing so, we can

[3] It should be noted that while this componential characterisation of control-related functions is something of a consensus in the neurocognitive literature, there is considerable debate as to the specific nature of these functions and while the components outlined here are the most common, it is by no means an exhaustive treatment.

begin to deconstruct these somewhat inscrutable executive mechanisms and so extend upon the foundations that cognitive science has laid.

1.4 Neural Network Representations as an Exploratory Tool for Cognitive Control

In the neural network account, as opposed to traditional perspectives on control, there is no need for reflection on internal stores of information to employ in some independent control mechanism. In these models, the only information retained is how the information should be processed. Connection weights merely describe how the signals shared between neurons should be passed on. Using such neurally plausible forms of representation means we can explore questions of control without deferring these functions to the homuncular mechanisms espoused in the traditional cognitive and philosophical literature. Examining questions of control then becomes a matter of determining how representations thus described can account for control-like behaviour.

Chapter 2. Control in the Absence of Control Representations

2.1 Motivation and Control

Given that cognitive control is often framed as the coordination of behaviour in service of a goal, we should first consider the animal's motivation to engage in purposive action. In his influential work on the principles of behaviour, Hull (1943, p. 17) identifies a guiding principle in the cognitive and behavioural sciences: "Since the publication by Charles Darwin on the Origin of Species it has been necessary to think of organisms against a background of organic evolution and to consider both organismic structure and function in terms of survival". On this view, to the extent that they have agency, organisms will be fundamentally motivated to perform actions which satisfy the dominant physiological need at any given point in time. An important corollary concerns the selective pressures brought to bear by the environment in which the organism subsists; the conditions which define survival vary as a function of the organism's ecological niche. Thus, organisms will act to attain, or maintain states of the world that promote their survival, and avoid that which might obstruct those states. Perception subserves this function of organisms, constituting the manner(s) in which an organism is sensitive to its environment and prompting it to act (or not act) in response (Allport, 1955; Mather, 2016; Pomerantz, 2003). Perception in organisms is not arbitrary, but has evolved to reflect the meaning of a given object or event to the organism; its *valence,* or its potential to contribute or detract from certain desirable states of the world (Barrett, 2006; Carruthers, in press; Lebrecht, Bar, Barrett, & Tarr, 2012), and the consequent possibilities of action afforded to the organism in the context (Feldman, 2016; Gibson, 1979; Millikan, 1995; Parker & Newsome, 1998).

This principle forms the basis of behavioural coordination. Even single-celled organisms must at times integrate chemosensory, sensorimotor, communicative and physiological information to inform adaptive responding (Lyon, 2015). To illustrate, consider

the chemotaxic properties of these creatures. Bacteria possess in their cell membrane molecules sensitive to the presence of glucose (a food) and phenol (a toxin). The presence of these substances automatically instigates a series of molecular feedback mechanisms that determine whether the bacterium's flagella movements are coordinated such that it moves up or down the encountered concentration gradient and so away from, or toward the substance (Adler, 1966; Alon, Surette, Barkai, & Leibler, 1999; Brandman & Meyer, 2008; Macnab & Koshland, 1972; Wadhams & Armitage, 2004).

This differential coordination of response highlights an important feature of organisms. Even at the lowest levels, organisms are capable of integrating information from competing signalling pathways to achieve adaptive outcomes (Hazelbauer, Falke, & Parkinson, 2008).

Parallels can be drawn in cases where flowers move in order to maximise their position in relation to the sun, known as phototropism. While in some cases the property appears to be purely mechanical, phototropism can be quite complex, involving multiple signalling pathways, photoreceptors, and hormones to coordinate differential growth gradients (Whippo & Hangarter, 2006).

In all cases, organisms exhibit the capacity to flexibly respond to environmental cues in order to promote survival, and often this requires the integration of signals to achieve differential outcomes. In the following section, I will argue that this consideration should feature heavily in an appreciation of cognitive control.

2.2 Behaviour Control is Realised in the Brain for More Sophisticated Animals

Godfrey-Smith (1998, 2002) advocated the notion that the complexity of an organism's ecological niche is the catalyst for the evolution of more complex forms of cognition. While such a hypothesis is debatable, it is certainly the case that the stimulus-

response propensities evident in unsophisticated organisms are eventually realised in the brain for more sophisticated animals (Allman & Martin, 1999; Damasio, 1999; Greenspan, 2007). Indeed, the work on adaptive integration of sensory signalling pathways in simple organisms in many ways underpins information-processing perspectives on cognition (Bray, 2009; Miller, 2003; Wadhams & Armitage, 2004). Recall that associative learning is characterised by changes in neural structure that elicit behaviours in response to stimuli that correspond to predictable outcomes. A very basic property of neurons is the strengthening of connections between cells that fire in synchrony, and the weakening of connections for those which fire at dissimilar intervals, commonly known respectively as 'Hebbian' or 'anti-Hebbian' learning (Bienenstock et al., 1982; Hebb, 1949). In this way, the brain marries perceptions with the appropriate responses for organisms, allowing for behaviour in environmental contexts which exceed the capacity for less complex biological mechanisms (Miller & Cohen, 2001). As such, where control in less sophisticated organisms is achieved without a brain, it seems sensible to assume a hierarchy of simplistic control mechanisms that are eventually realised in the neural mechanisms of more sophisticated animals.

Recall that the information contained in a neural network is how the signal transduced by input neurons should be passed on to the response neurons; the perception of a stimulus automatically triggers the associated response. Neural plasticity can account for the development of these stimulus-response relationships in the brain. However, it is not sufficient to explain control over those relationships which compete, merely to account for which stimuli are linked to which responses. Augmenting an account of neural plasticity with a neurally plausible account of network function provides insight into how control might be achieved.

2.3 Evidence-Accumulator Models as a Foundation for Control

A family of computational models known collectively as *accumulator models*, feature competing neural network representations as evidence-accumulators. These accumulators each independently gather channels of appropriate perceptual information. When one channel reaches some threshold, it will 'win out' over the others in specifying a response (e.g. Cisek, 2007; Ditterich, Mazurek, & Shadlen, 2003; Ratcliff, 1978; Usher & McClelland, 2001; Vickers, 1970; Wang, 2002). The form of this evidence is taken to be the overall activation of a given neural network representation. This 'winner-take-all' property in such models is particularly attractive because it can account for circumstances in which individual neural network representations share similar inputs, or similar outputs; only the most active representation will be fully realised.

An important theoretical development in these models has been the introduction of reciprocal inhibition, in which all related representations inhibit the others proportional to the amount of appropriate information obtained (Fig. 3; Basten, Biele, Heekeren, & Fiebach, 2010; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Feng, 2012; Krajbich & Rangel, 2011; Marshall, Bogacz, & Gilchrist, 2012). This addition has been crucial in aligning the computational models with the performance outcomes achieved in real-world experimental data for a range of psychophysical tasks (e.g. Laming, 1979; Ratcliff, 1978; Stone, 1960). These models match, for example, the real-world data on decision accuracy and reaction time distributions (Brody & Hanks, 2016). Likewise, accumulator models approximate the underlying neural processes well, accounting for neural activation patterns related to perceptual choice tasks (Gold & Shadlen, 2007; Schall, 2001; Shadlen & Newsome, 2001; Usher & McClelland, 2001). For example, the rate at which the evidence accumulates toward a decision threshold in these models is comparable to the decision-relevant neural response variability measured in primate brains during perceptual tasks (Gold & Shadlen, 2007).

Further, while the vast majority of experimental evidence pertains only to decisions involving two alternatives, evidence-accumulator models can be extended to incorporate multiple alternatives (Krajbich & Rangel, 2011; Usher & McClelland, 2001) and multiple steps of action (Solway & Botvinick, 2015).

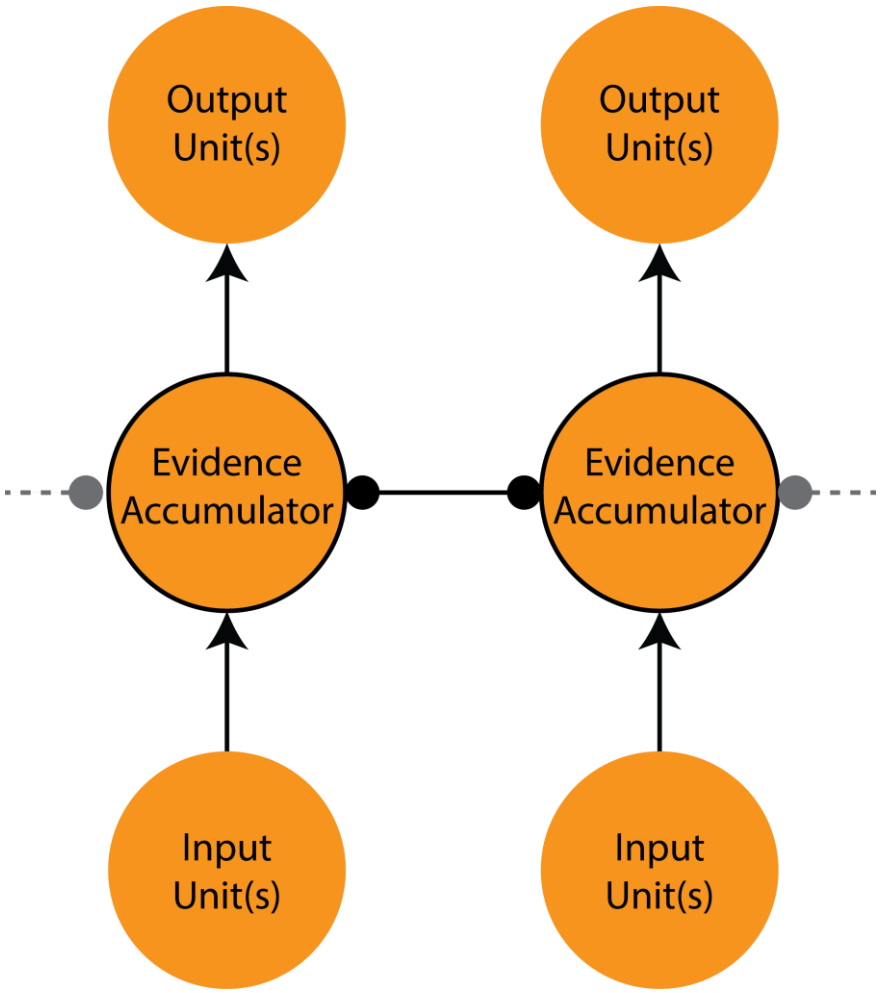

Figure 3. Diagram of an evidence accumulator model. Arrows indicate excitatory connections. Closed circles indicate inhibitory connections. Signals from input units converge on an evidence-accumulator for each neural network representation. Lateral inhibitory connections inhibit competitors proportional to the activation of respective input units. When a specified decision threshold of activation has been reached, one representation will 'win-out' over the others in passing the signal on to output units determining the response.

2.3.1 Evolutionary valence can inform action-selection. These accumulator models can explain how the connection weights of real world neural network representations are changed to achieve control with reference to their current interactions. In the models, the combination of accumulation (equivalent to neural excitation) and the inhibitive capability of

the links between these pathways can be used to model the optimisation of resource allocation demonstrated in neural systems. The experimenter will commence by specifying the initial drift rate (equivalent to baseline neural activity), or the decision threshold for competing representations.[4] The parameters defining these evidence-accumulators (i.e. the criteria for triggering a threshold) will then change according to an algorithm designed to approximate the plastic properties of neurons. This algorithm will seek to achieve the optimal compromise between the speed of evidence accumulation and accuracy (Bogacz et al., 2006; Marshall et al., 2012). For example, a threshold requiring less evidence to trigger would be faster, but sacrifice accuracy in action-selection. Given an acceptable error rate,[5] this may explain the manner in which real world neural network representations governed by the plastic properties of neurons would establish naturalistic dominance patterns over time and development. In doing so, evidence accumulators achieve the most basic functions of control; the selection of appropriate responses, and the inhibition of inappropriate ones.

Arguably, these mechanisms also satisfy functions typically ascribed to attentional processes. Attention tends to be vaguely defined, but is generally considered to be a process responsible for tuning the parameters of information processing in order to optimise task performance (Chun, Golomb, & Turk-Browne, 2011; Tsotsos, 2017; see also section 2.6.2). In the cognitive literature, it has become commonplace to assume that the executive performs attentional functions. This is largely due to the observation that some effortful attention is employed in executing tasks such as the Stroop (recall section 1.3).

[4] It should be noted that neural networks have seen myriad applications in tasks that vary wildly in their ecological verisimilitude. Here, I concern myself only with experimental intervention as informed by actual neural function, and propose that in organisms, these would be established by the relevant evolutionary imperatives.

[5] Again, an experimenter defined parameter, assumed to be established organically in an animal (e.g. I imagine mice inherit a high acceptable error rate for a fear response in the presence of a possible threat, it being far less adaptive to ignore it).

However, there is no requirement to invoke an independent attentional process to explain the optimisation occurring in these evidence-accumulation processes. Rather, it is intrinsic to their function. So long as the decision threshold has been specified, the parameters will be adjusted automatically to optimise performance for speed and accuracy in action-selection.

Numerous networks which exhibit these evidence-accumulation and reciprocally inhibitive properties can be found in the brains of both vertebrates and invertebrates (Barron, Gurney, Meah, Vasilaki, & Marshall, 2015; Brody & Hanks, 2016; Gallistel, 2013; Windhorst, 1996). As such, mechanisms that can achieve control over low-level environmental conflicts surely exist in the absence of an executive mechanism.

Yet, while a great number of the actions an organism engages in are motivated by the evolutionarily implied valence of stimuli (see Panksepp, 2004 for a treatment),[6] a great many more are informed by experiential interactions with the world. Learning is a crucial feature of adaptive behaviour and to account for this in the models just described, we require further mechanisms with which to incorporate feedback.

2.4 Learning Informs Value-Based Action Selection

It is fortunate then, that incorporating learned valence into the existing evidence-accumulator models poses no significant challenge. Further, by examining biological mechanisms of learning, we can infer the real-world processes responsible for altering the defining parameters of accumulator mechanisms to adapt to a changing environment.

Learning valence is a phylogenetically basic capacity (Lyon, 2015; Skinner, 1938; Thorndike, 1911). Very simple animals can learn new possibilities for the potential of stimuli

[6] Although any taxonomy of these will be largely speculative, given that one cannot empirically deduce the full evolutionary history of an organism.

to contribute or detract from desirable states of the world (Alem et al., 2016; Menzel, 2013; Webb, 2004). The significance of this observation should not be underplayed. It is difficult to imagine many ecological scenarios in which the kinds of speed and accuracy judgements described above solely inform action-selection.[7] Rather, there are many environmental variables and many distinct actions available to an animal which must be judged for their relative utility in any given circumstance. However, to make a utility judgement, the organism must possess information about the outcomes of an action in a given environmental context and be able to determine how that might relate to its goal (Pirrone, Stafford, & Marshall, 2014).

2.4.1 The value of expected outcomes may be judged for their contribution to a goal. Cognitive science typically assumes that this information is stored in some propositional form and reflected upon to inform goal-directed behaviour (section 1.1; 1.3; Buckner, 2010; Damasio, 1999; Doya, 2008; Seligman, Railton, Baumeister, & Sripada, 2013; Wilson & Gilbert, 2005). In this vein, evidence-accumulator models have been extended to make utility-based judgements, using the same kinds of reciprocal inhibitory connections described earlier (Basten et al., 2010; Krajbich & Rangel, 2011; Pais et al., 2013; Rangel & Hare, 2010; Rustichini, 2008; Seeley et al., 2012; Usher, Elhalal, & McClelland, 2008). However, this is typically done by incorporating an independent representation of the extrinsic value of stimulus features to inform competitive decision making. This approach does not entirely release us from the homuncular problem. We must still explain the mechanism responsible for determining the value of stimuli according to the expectation that they might facilitate the animal's goals. However, there may be no need to posit such a

[7] Indeed, only one comes to mind; imminent life or death scenarios, in which a fast but inaccurate decision would likely prove more beneficial than slower, more accurate ones (e.g. Trimmer et al., 2008).

mechanism in all cases (Gold & Shadlen, 2007). A more plausible interpretation of utility-based judgements considers value as an intrinsic property of stimuli.

2.4.2 Value is an intrinsic property of neural network representations. In the presence of reinforcement (reward or punishment), any organism with the capacity for learning develops an extensive repertoire of outcome-expectations in service of adaptive function (Lyon, 2015; Menzel, 2013; Rescorla & Solomon, 1967; Skinner, 1938; Thorndike, 1911; Webb, 2004). These expectations merely comprise the valence of a stimulus-response association in a given environmental context. When hungry, the valence of food becomes salient, the animal having learned that the food item will fulfil the need. When satiated, food no longer presents as salient, and the animal may concentrate on some other dominant need.

Observe that this requires no explicit representation of a goal, nor an independent representation of an expectation to inform the animal of the stimulus' goal-related value. Instead, they can simply be viewed as neural predispositions to respond in certain ecological circumstances (Carruthers, in press; Dreyfus, 2002; Frijda, 2016; Gollwitzer, 1999; Jeannerod, 2006; Pacherie, 2002; Scott, 2006). This feature of reinforcement learning should not be trivialised. The assumption that goals are explicit has encouraged lines of enquiry devoted to discovering the mechanisms responsible for selecting among them (e.g. Ravizza & Carter, 2008). While explicit goals may be realised in the brains of higher-animals, there is no reason to accept that all goal-directed behaviour is accomplished thus. Similarly, there is no need to assume that the expected outcomes of an action are explicitly measured against a goal-state.

Indeed, expectations can be explained in the context of neural network representations, the premise of which being that environmental contexts activate inner states which intrinsically afford certain responses to the organism. In such cases, the relative utility of a representation is an intrinsic property (Carruthers, in press; Peil, 2014) that can be used

as evidence in a more generic reciprocally inhibited network (Basten et al., 2010; Bogacz et al., 2006; Pais et al., 2013; Rangel & Hare, 2010). This development has permitted the extension of accumulator models to account for circumstances in which a decision must be made between alternatives which have equal, or equally uncertain values (Pais et al., 2013; Pirrone et al., 2014). Thus, by remaining close to the fundamental principles of reinforcement learning, we avert the risk of invoking homuncular mechanisms where none need be present.

Examining the features of reinforcement learning not only permits insight into how animals make utility judgements, but also how these judgements can be optimised over time.

2.4.3 Biological mechanisms of reinforcement learning provide insight into the how parameters are altered. In a reciprocally inhibited network architecture, the optimisation of the system depends on three key parameters; the baseline activation of the neural network representations, the rate of evidence accumulation (in the form of neural response variability), and the threshold for triggering a response. In the brain, reinforcement learning is primarily facilitated by the interaction between changes in neuromodulator concentrations across neuronal populations and the Hebbian and anti-Hebbian properties of neurons, contributing to the growth and stability of connections (Doya, 2008; Soltoggio, Durr, Mattiussi, & Floreano, 2007). Neuromodulators are associated with the neural equivalents of each of the parameters defining evidence-accumulator models. Neural response variability has long been observed to be influenced by the presence of nor-adrenaline (Keeler, Pichler, & Ross, 1989; McGinley, David, & McCormick, 2015; Solway & Botvinick, 2012) and there is evidence to suggest that acetylcholine provides information regarding the certainty of stimuli, a factor that drives the rate of accumulation (Angela & Dayan, 2005; Marshall et al., 2016; Sarter & Bruno, 1997). Dopamine (e.g. Braver & Barch,

2002; Niv, 2009) and oxytocin (e.g. Kis, Hernádi, Kanizsár, Gácsi, & Topál, 2015) have been known to influence the expectancy biases observed in humans and other animals and could serve to alter the baseline activation of neural network representations.

The final relationship deserves some special attention, being both well-studied and bearing a striking similarity to the performance monitoring function typically ascribed to some executive process in classical cognitive science. Unexpected events, or *prediction errors*, are presumed to play a major role in reinforcement learning; unexpected outcomes may indicate that the organism's environment has changed, and signal the need for adaptation (Hohwy, 2013; Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1990). Work conducted by Wolfram Schultz (1997) formalised this notion, demonstrating that unanticipated rewards in vertebrates are signalled by phasic dopamine responses. Electrophysiological data routinely identifies neurons in mammals that signal reward-related information in the mid-brain dopaminergic pathways that are suitable candidates for learning and decision-making (Bromberg-Martin, Matsumoto, & Hikosaka, 2010; Schultz, 2016). The same can be said of octopamine and dopamine in the invertebrate brain, which signal aversive and appetitive stimuli respectively and may modulate connections accordingly (Perry & Barron, 2013; Søvik, Perry, & Barron, 2015). Indeed, the incorporation of octopamine and dopamine in the simulation of a partial insect brain has been used to successfully model reward and punishment learning in insects (Bazhenov, Huerta, & Smith, 2013; Schwaerzel et al., 2003; Vergoz, Roussel, Sandoz, & Giurfa, 2007). Based on these findings, many neural network models incorporate prediction error as a means of adjusting decision thresholds. Different threshold values are sampled and algorithmically adjusted to optimise for the ideal rate of reward. Assuming that neuromodulatory changes take the place of these algorithms in organisms, the values of these decision-making thresholds could conceivably be adjusted by an interaction between these neuromodulator concentrations and the disposition of the

connections in the neural network architecture. Recall that in the classical literature, performance monitoring is typically ascribed to cortical processes (Petrides & Milner, 1982) and is traditionally assumed to be carried out by a supervisory mechanism (e.g. Norman & Shallice, 1980). On this view, some independent arbitrator presides over lower-order processes to iteratively improve their performance. In stark contrast, evidence accumulator architectures provide convincing evidence that such a feature is, at least in part, an emergent property of a reciprocally connected neural network. There is no requirement for an external arbitrator, nor a need for cortical intervention. Performance monitoring in these models is simply the naturalistic outcome of phasic neuromodulator responses.

As such, incorporating neuromodulatory feedback into the existing evidence accumulation models theoretically permits an organism to optimise action-selection according to speed, accuracy, and relative utility as an inherent property of neural network function (Bogacz & Gurney, 2007; Dayan & Daw, 2008). Evidence-accumulator models then, present a neurally plausible mechanism which may account for the resolution of a wide range of low level conflicts. However, at more ecologically complex levels, these models betray important limitations.

2.4.4 Accumulator models will fail in more complex environments. With a suitably complex organism, the scope of relevant ecological considerations will eventually outstrip the capacity of simple evidence-accumulator models to explain.

The optimisation of the systems described in the preceding section would only occur in the presence of events which, through reward or punishment, trigger the activity of certain neuromodulators. Neuromodulatory changes tend to be quite rapid, and these architectures do not therefore lend themselves well to circumstances involving delayed feedback. Delayed consequences are associated with significant decrements in performance during operant

learning paradigms (see Lattal, 2010 for a review), hinting at the fundamental role of neuromodulator action during learning. Yet, sophisticated animals clearly at times perform in the absence of immediate feedback.[8] A well-known example in this regard concerns the ability of rats to apparently learn the organisation of a maze prior to the introduction of any reinforcement (Anthony, 1959; Seward, 1949; Tolman & Honzik, 1930; Wirsig & Grill, 1982).

Further, ecological scenarios tend to provide more information to an organism than merely the potential for reward or punishment. Suitably sophisticated organisms might use this information to inform action-selection in other ways, by assessing the opportunity costs involved for example, or the difference in energy expenditure (Frank & Claus, 2006; Gopnik, Schulz, & Schulz, 2007; Niv, 2009). While these factors have been explored from an optimisation perspective (in a fashion similar to that described in 2.4.1; e.g. Aston-Jones & Cohen, 2005; Yu & Dayan, 2005), they do not yet map neatly to plausible neural mechanisms.

It must also be noted that with a large number of competitive representations, these architectures would become prohibitively dense. The crucial feature of accumulator models in aligning the model data with real-world outcomes is the addition of reciprocal inhibition. To 'win out', a representation must inhibit its competitors proportional to the amount of evidence it has accumulated, a process requiring inhibitory links between each related representation. For each new competitor into the system we must therefore introduce reciprocally inhibitory links to each existing competitor, increasing the wiring exponentially

[8] There do exist special reflexive mechanisms for this, although these would not account for all circumstances in which delayed consequences inform learning. Consider, for example, the phenomenon of taste aversion (Bures, Bermúdez-Rattoni, & Yamamoto, 1998; Garcia, Ervin, & Koelling, 1966). With minimal exposure, a substance which induces illness will result in a long-term aversion to said substance, despite significant delays between ingestion and the onset of symptoms. Taste aversion may in fact belong to a wide range of similar reflexive associations, which have been brought under the heading of preparedness learning (Seligman, 1971).

to account for every possible conflict (Lennie, 2003; Redgrave, Prescott, & Gurney, 1999). It seems likely that the metabolic cost of operation would eventually outweigh the adaptive benefit of such an architecture, placing pressure on solutions that more efficaciously achieve comparable performance. This would be compounded by the fact that, as an organism and its environment become more sophisticated, so too would the value relationships between stimuli. In a reciprocally inhibited network, in which all competitive representations are linked, modifying one representation to better account for its comparative utility to another may sub-optimally reflect its comparative utility in other circumstances. For example, on learning that yellow flowers are less nutritious than red, a bee may inherit a slower accumulation rate for yellow. Later learning that red flowers are less nutritious than green flowers, red may suffer the same consequence, invalidating the earlier utility distinction between red and yellow. Indeed, valence can be produced from many properties of stimuli depending on the circumstance, making overall evaluative responses increasingly difficult in a reciprocal network alone.

2.5 Centralised Action-Selection

A more metabolically efficient option would be a centralised selection mechanism, in which neural network representations would converge on a single locus, allowing the conflict between competitors to be evaluated and resolved therein. Such a system requires only two connections for each representation; one input to and one output from the central mechanism (Barron et al., 2015; Redgrave et al., 1999).

In vertebrates, such a mechanism is thought to be realised in the basal ganglia (BG) and associated neural structures, which both collects input from diverse neural areas across the cortex and sub-cortical regions (McHaffie, Stanford, Stein, Coizet, & Redgrave, 2005), and contains the requisite properties to support the competitive processing of neural

network representations (Gurney, Prescott, & Redgrave, 2001; Redgrave et al., 1999). On this hypothesis, inputs, termed *action requests*,[9] comprise the overall activity of a given representation, signalling the salience of a neural activation. These action requests are then subjected to competitive processing as they travel through *action channels*[10] within the BG to determine which is most salient. The BG 'selects' this input by decreasing the supply of inhibition in the corresponding action channel (while maintaining, or perhaps increasing the inhibition of non-selected channels). This permits the most salient representation to 'win-out' while inhibiting the competitors. When the relevant features of the BG are applied in a computational model, the simulations achieve a Bayesian-like optimisation of conflict resolution (Bogacz & Gurney, 2007; Lepora & Gurney, 2012).[11] This suggests that the BG system does indeed possess the requisite features to plausibly explain the optimisation of performance in an ecological scenario.

The primary benefit of the proposed structure is the use of salience as a 'common currency', allowing for complex valence relations to produce an overall evaluative response when competing alternatives would otherwise appear incommensurable (Carruthers, in press; Gurney et al., 2001; Levy & Glimcher, 2012; Redgrave et al., 1999). Such a mechanism would have the added advantage of modifying individual representations without necessarily altering the functionality of others.

[9] The proponents of this hypothesis use the term 'action request' as something of an illustrative catch-all for system inputs, failing to specify their exact nature. However, if we consider these to be neural network representations, the hypothesis carries more weight.

[10] Consider these to be anatomically discrete populations of neurons associated with distinct action requests, described best in Bogacz and Gurney (2007); Gurney et al. (2001). Again, we can view these as neural network representations whose inputs are not external stimuli, but instead the neural network representations comprising an action request.

[11] Bayesian optimisation is a strategy frequently employed by modellers to simulate the combination of new evidence with prior beliefs or expectations in a principled manner (via the application of Bayes' rule) (Mockus, 2012; see also section 3.3.2).

Both the basal ganglia and the striatum are known to possess the requisite features to support a cross-inhibitory architecture in a number of configurations, and a similar functionality has been proposed for invertebrates in the lateral protocerebrum (see Barron et al., 2015 for a review). Indeed, the latter was shown to be capable of theoretically achieving comparative outcomes with the vertebrate system just described despite non-trivial system differences.[12] The key outcome of this exploratory work indicates that a system architecture supporting centralised processing of competing representations is a plausible neural specialisation for increasingly sophisticated organisms to develop in order to achieve control in increasingly complex environments.

2.6 Executive Functioning with No Executive

We entered this chapter asking how the traditional components thought to comprise the executive might be explained in a neurally plausible manner. In the interest of clarity, I will now address each and highlight again the opportunities described in the preceding sections for these functions to be achieved as a feature of neural network dynamics.

2.6.1 Appropriate action-selection and inhibition. Adaptive action selection is a problem that faces even single-celled organisms, and the capacity to integrate competing signals in service of this precedes the development of neural tissue (Bray, 2009; Hazelbauer et al., 2008). That the resolution of conflicts at the level of the brain necessitates in all cases homuncular arbitrators, or top-down 'executive' control thus seems unlikely. Indeed, a reciprocally inhibitive neural network architecture supports a vast array of self-organising decision-making processes and maps well onto the existing neuroanatomical structures found in both vertebrate and invertebrate species. I do not make the claim that such mechanisms are sufficient to resolve conflicts in all ecological contexts. However, that these networks are

[12] Albeit on a less complex scale.

capable of optimisation in service of speed, accuracy, and relative value utilising only known properties of neural function should raise questions about the continued utility of invoking homuncular mechanisms to account for control-related phenomena.[13]

2.6.2 The role of attention in control. As animals develop in sophistication, so too will they develop perceptive capabilities that increasingly exceed those strictly required for immediate tasks. As such, attentional processes are considered a vital component of cognitive control phenomena. While the facets of attentional processes have provoked myriad theoretical treatments (see Chun et al., 2011), the most enduring characterisations involve the ability of organisms to adjust input and computational load to that best suited to achieving the task or tasks at hand.[14] In computational terms, attention is best viewed in terms of a dynamic tuning of processing in response to task demand; attentional processes adaptively tune the parameters of information processing in order to optimise task performance (Tsotsos, 2017). Traditional approaches to the question of attention assume that an organism requires a set of established representations in order to perform this function, and these are activated to the appropriate degree when required (e.g. Botvinick et al., 2001; Norman & Shallice, 1980). Yet, thus far, there has been no need to invoke such representations. Provided the system possesses the requisite decision thresholds, here proposed to be established and modified by selective pressures or learned valence, this parameterisation is an emergent property of the network. Both baseline activation and

[13] While the focus of this thesis is on the resolution of conflict in service of control, the application of neural network representations can assist to narrow enquiry into other, more amorphous aspects of control also. For example, Rodney Brooks' (e.g. 1990, 1999) *subsumption architecture* bears similarities to biological nervous systems and demonstrates how some representations may impose control on less dominant representations in limited ecological contexts (Gallistel, 2013; Maes, 1993; Prescott, Redgrave, & Gurney, 1999; Redgrave et al., 1999).

[14] Often, a distinction is made between attentional processes responsible for selecting among competing alternatives (e.g. Desimone & Duncan, 1995) and attending to a selection to improve performance (e.g. Carrasco, Ling, & Read, 2004). I make no such distinction here, in part because doing so is beyond the scope of this thesis, but also because in many cases this distinction is unclear. Performance-enhancing attentional processes often involve competitive processing internally (see e.g. section 3.5.2).

accumulation rate can be adjusted by the associative properties of neural function to achieve the optimal compromise between speed to accuracy, permitting adaptive performance across a broad range of contexts (Bogacz et al., 2006; Marshall et al., 2012; Pais et al., 2013; Rangel & Hare, 2010; Usher & McClelland, 2001).

2.6.3 Performance monitoring with no monitor. Recall that a fundamental attribute of reciprocally inhibitive architectures are the thresholds which trigger a decision, allowing one representation to 'win-out' over others. Many models incorporate prediction error as a means of adjusting the value of these thresholds; different threshold values are sampled and adjusted according to an algorithm to optimise for the ideal rate of reward. This property of neural networks bears conspicuous similarity to the phenomena of 'performance monitoring' that is typically ascribed to cortical executive processes in the cognitive literature (Norman & Shallice, 1980; Petrides & Milner, 1982). Here, some supervisory process is thought to adjust lower-order processes in order to improve an animal's performance in a given task. However, given that unexpected events, or prediction errors, are signalled in vertebrates by phasic dopamine responses (Bromberg-Martin et al., 2010; Schultz, 2016; Schultz et al., 1997; Sutton & Barto, 1990) and in invertebrates by octopamine and dopamine (Bazhenov et al., 2013; Perry & Barron, 2013; Schwaerzel et al., 2003; Søvik et al., 2015; Vergoz et al., 2007), the values of these decision making thresholds may instead be adjusted by the change in neuromodulator concentration across the connections of the relevant evidence-accumulators. That is not to say that all performance monitoring can be accounted for by prediction error. In fact, prediction errors are demonstrably inferior mechanisms for regulation, having only a limited capacity to resolve conflicts, and higher organisms evolve progressively more effective mechanisms to inform performance improvements (Conant & Ashby, 1970).

2.6.4 Goal-directed behaviour does not require goal representations. The development of expectations, in the form of behaviour-outcome contingencies, is a phylogenetically basic property of organisms. In the learning literature, this is commonly referred to as 'instrumental' or 'operant' learning. Commencing with the ostensible birth of the domain of study (Skinner, 1938; Thorndike, 1911), accounts of goal-directed behaviour typically make a distinction between behaviour elicited by some triggering stimulus, and that which is controlled by the animal's knowledge of the potential consequences of an activity (although c.f. Rescorla & Solomon, 1967). This distinction has given rise to a widespread notion that instrumental behaviour should be explained by the storage of expectations in some propositional form to be operated on in service of some intention. However, there is no reason to conclude that all such goal-directed behaviour is accomplished thus. The brains of even quite simple animals contain information regarding the dependencies of anticipated states on earlier states and actions (Menzel, 2013; Webb, 2004). Such expectations can be plausibly explained in the form of neural network representations, in that the environmental context activates a representation which intrinsically affords certain responses to the organism. Given the capacity to learn outcome-expectations, an animal needs no explicit representation of a goal or intention to engage in purposive behaviour, and control over conflicting priorities can be incorporated into the existing evidence-accumulator models as the difference between the intrinsic, or learned valence of expected outcomes.

2.6.5 Working memory. Conspicuously absent from the sections above is any discussion of the holding or integrating information over time (i.e. working memory). Evidence-accumulator models are quite limited in this regard. Recall that the optimisation of reciprocally inhibitive architectures is largely dependent on neuromodulatory changes across neuronal populations (section 2.4.4). Such changes tend to be rapid, and these architectures

do not therefore lend themselves well to circumstances involving delayed feedback.[15] This limitation does not necessarily restrict the function of evidence-accumulators, however. Contemporary accounts of working memory rely primarily on reverberatory, localized persistent activity (Compte, Brunel, Goldman-Rakic, & Wang, 2000; Freeman, 1995; Goldman-Rakic, 1995). Reverberatory systems such as these can be found at any level of the brain, and incorporation of this widespread property of neural function into the existing models of lateral-inhibition may provide insight into how such systems may hold and integrate information over time.

2.7 Conclusion: Moving Toward Control Representations

The preceding sections highlight the opportunities available for the brain to achieve control in the absence of higher-order control representations or cortical control mechanisms. An account of behavioural control in the brain can be largely explained by way of the adaptive tuning of neural network representations to attain, or maintain desirable states of the world and avoid that which might obstruct those states. The signals indicating that control is necessary may be evoked by the evolutionarily-hardwired, or experientially-learned valences of stimuli. More significantly, the brain can support control-related phenomena typically ascribed to 'executive' functioning, without any homuncular modulation of 'lower-order' processes.

Yet, such systems cannot account for all control-related phenomena. More complicated animals engage in behaviours that frankly outstrip the capacity for the models we have described. To adaptively resolve conflicts, evidence accumulators are reliant on highly context sensitive representations. This context-sensitivity allows the mechanism to integrate the difference in available evidence between competing representations. However,

[15] A notable exception in this regard can be found in Pais et al. (2013).

keeping with Godfrey-Smith's (1998, 2002) thesis, as an organism moves toward increasing levels of sophistication, the brain must engage in more flexible ways to represent wider ranges of behavioural complexity in response to a wider range of environmental considerations (see Lazareva & Wasserman, 2008; Seger & Miller, 2010; Taylor & Stone, 2009; Zentall, Wasserman, & Urcuioli, 2014). Let us then explore the ways in which neural network function might help us explain more complex behaviour.

Chapter 3: The Anatomy of a Control Representation

3.1 Domain-General Representations as a Control Mechanism

In an increasingly complex world, a brain requiring representations which account for each individual pairing of environmental context to action set would eventually place a prohibitive burden on the system. Since Lashley (1951), it has been generally accepted that sophisticated organisms, with particular reference to primate species, are capable of detecting the commonalities shared across experiences and thus group them into meaningful clusters to coordinate behaviour in complex environments. As such, where control is initially achieved by highly context sensitive representations, more domain-general representations would likely develop to exploit the naturalistic properties shared across environmental contexts (Botvinick & Cohen, 2014; Karmiloff-Smith, 1995; Lake, Ullman, Tenenbaum, & Gershman, 2016). These would permit an animal to more flexibly apply acquired knowledge in new ways and coordinate behaviour in new contexts, allowing animals to generalise performance across tasks sharing environmental commonalities.

We are thus left to question how. Botvinick and Cohen call this a "metaoptimization problem":

> *"Given a particular distribution of naturally occurring tasks, how can the control system itself be configured so as to optimize performance across tasks? The objective function here involves not only single-task performance but also the generalizability of control—that is, the efficiency ("economy of scale") gained by using similar representations to control multiple tasks"* (2014, p. 1262).

To highlight the necessity of an account of 'metaoptimisation' as it relates to control, let us return to the example of the Stroop task (Fig. 1). Cognitive demands may come into conflict for a number of reasons, depending on the nature of the task parameters (Cohen et

al., 1990; Engle, 2002; Kahneman & Treisman, 1984; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). It is likely, for example, that much of the difficulty experienced in executing the Stroop is evoked by the need for a more practiced, or 'automatic' process (word reading) to be overcome by a less commonly employed one (colour naming) (MacLeod & Dunbar, 1988; Posner & Snyder, 1975).

Yet, difficulties like these would not emerge unless induced by the architecture of the brain. A fundamental challenge to any networked system is that of *mutual interference*, or *cross talk*. Cross talk occurs when two tasks share local resources, and so compete for their use (Allport, 1955; Cohen et al., 1990). In the Stroop, it would appear that the neural processes for both colour naming and word reading share systems (for example, those required to form a verbal response), and one must overcome the resultant cross talk to complete the task (Cohen et al., 1990; Feng, Schwemmer, Gershman, & Cohen, 2014).

As the nature of tasks in which an organism engages becomes more complex, the likelihood that these tasks will share common processes increases, making their simultaneous performance impossible where these shared resources are integral to their execution (Navon & Gopher, 1979; Allport, 1980; Allport, Antonis, & Reynolds, 1972; Logan, 1985; Wickens, 1984). Neural network models apply this at the level of the neuron; this cross talk arises when tasks recruit neural network representations that share neurons responsible for encoding input and specifying responses. This poses a control problem when one representation must be selected over another (recall section 2.3). Similarly, when a representation is recruited in service of multiple different tasks, or *multiplexed*, the brain will face difficulty in prosecuting those tasks simultaneously (Fig. 4; Feng et al., 2014).

3.1.1 Task 'rules' inform conflict resolution. To resolve this particular conflict, some additional factor must intervene to bias action selection in favour of one response or

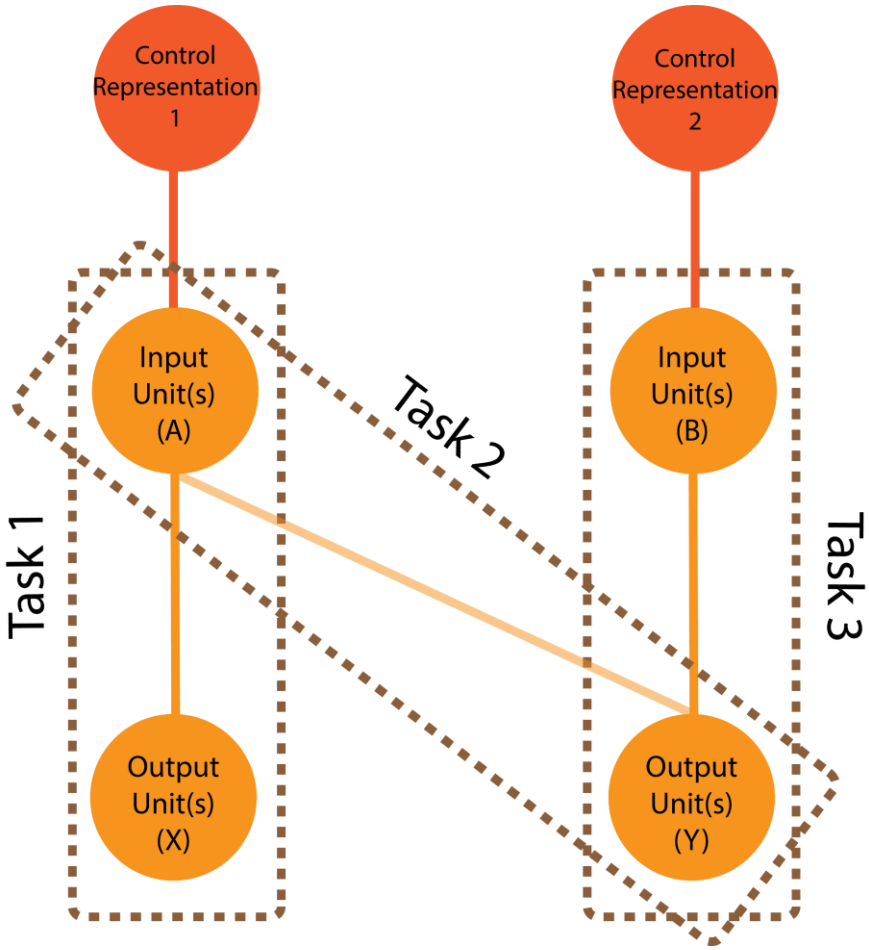

Figure 4. Multiplexing of multiple tasks. This diagram illustrates three tasks (brown dashed lines) which engage multiplexed representations (diagonal connection represents multiplexing). Tasks 1 and 2 share input units (A). Tasks 2 and 3 share output units (Y). Carrying out Tasks 1 and 2, or Tasks 2 and 3 simultaneously would result in conflict as they compete for shared resources. The connection between A and Y is less dominant (indicated by light colour) than the connection between B and Y. Where Tasks 2 and 3 came into conflict, representation B-Y would win out without training. For A-Y to win out in a transient context, Control Representation 1 would bias A-Y to increase its responding allowing it to overcome the prepotent B-Y connection.

another. This poses no significant challenge alone. As has been discussed (sections 2.3.1; 2.6.2), while this may be accomplished by some separate attentional or executive function, it need not be; evidence accumulator models describe an example of how this might be achieved in-house, so to speak. Typically, evidence accumulator models are used to resolve conflicting stimuli rather than conflicting responses. However, neural network representations make no distinction between the two; the stimulus affords the possible responses. As such, where a single stimulus affords two competing responses, we can

imagine that an accumulation process may suffice to allow one to 'win out'. Yet, recall that the Stroop task is further complicated by the presence of automaticity—the participant must overcome the prepotent word-reading response to name the colours. While we might account for the initial bias in the networks described earlier and account for how this bias might change with training,[16] it is far more difficult to imagine how this bias is mitigated such that the weaker response 'wins out' in the short term in accord with the transient task demands of the Stroop.

Cohen et al. (1990) solve this issue in a traditional neural network model by introducing what they call 'task demand units' (Fig. 4). These are independent control representations which serve to bias processing in favour of the weaker representations in circumstances which demand this outcome. In the context of the Stroop task, these task demand units are imagined to represent the 'rule' determining the nature of the task at hand, and so bias the colour naming pathway to make it more responsive to its input (see also Cohen, Aston-Jones, & Gilzenrat, 2004, pp. 73-74). Cohen et al. (2004) suggest these task demand units are realised in the pre-frontal cortex, which are thought to sustain the activation of representations using the recurrent mutually excitatory connections characteristic of this brain region. The authors go on to make various plausible proposals for how these they may be 'switched on' using existing neural structures (Botvinick & Cohen, 2014 review these in more detail; see also section 3.5.5).

One is then left to question what forms such representations might take, and how these might contribute to control.

[16] Again, for example, consider that neuromodulation theoretically permits us to modify the baseline activation of neural network representations in an evidence accumulator model (section 2.4.3). This might explain the propensity to engage in one of these tasks over another. However, when one is solving the Stroop, it is not likely that one is altering baseline activation – to alter the prepotent response requires training (MacLeod & Dunbar, 1988).

3.2 Classical Approaches to Domain-General Representations: Rules, Categories, and Concepts

Traditional models of control place heavy emphasis on explicit domain-general representations to explain an animal's ability to discern the value of stimuli across contexts (Anderson, 2013; Cowan, 2001; Engle, 2002; Miller & Cohen, 2001). These representations are referred to quite interchangeably as rules, categories, and concepts (Laurence & Margolis, 1999; Zentall, Galizio, & Critchfield, 2002), but can be broadly defined as a "knowledge of groupings and patterns [of functional relevance] that are not explicit in the bottom-up sensory inputs" (Seger & Miller, 2010, p. 2). Based primarily on experimental observations of vertebrate behaviour, the classical literature presents the capacity to structure knowledge thus as a hierarchy of increasingly complex, discrete abilities (e.g. Herrnstein, 1990; Katz, Wright, & Bodily, 2007; Lazareva & Wasserman, 2008; Mackintosh, 2000; Thomas, 2012; Thompson & Oden, 2000; Zayan & Vauclair, 1998; Zentall et al., 2014). These range drastically from simple associative learning (for example discrimination based on a representation of colour) to abstract conceptual ability (e.g. discriminating more from less using some kind of numerical representation).

3.2.1 The homunculus unnecessarily rears its head again. Borne out of, and complementing this hierarchical approach to categorical ability, the predominant neurological perspective distinguishes between behaviour based solely on stimulus-response relationships, and that which is coordinated by some cortical or otherwise executive function according to a rule (Fantino, 2003; Miller et al., 2003). This is largely due to the fact that damage to cortical regions appears to impair categorical decision tasks, particularly rule-learning and rule-switching in primates and humans (see Owen, Sahakian, Semple, Polkey, & Robbins, 1995; Seger & Miller, 2010; Stuss et al., 2000) and the extraordinary task specificity evidenced in mammalian cortical cells (Asaad, Rainer, & Miller, 2000; Durstewitz, Vittoz, Floresco, &

Seamans, 2010; Everling, Tinsley, Gaffan, & Duncan, 2006; Hoshi, Shima, & Tanji, 1998; Karlsson, Tervo, & Karpova, 2012; Miller, Erickson, & Desimone, 1996; Miller et al., 2003; Schoenbaum & Setlow, 2001; Tsujimoto, Genovesio, & Wise, 2011; Wallis, Anderson, & Miller, 2001; Wallis & Miller, 2003; White & Wise, 1999). Correspondingly, researchers have sought to identify the executive mechanism responsible for this coordination, with an emphasis on prefrontal and other cortical areas, or their analogues in non-mammalian animals (Güntürkün, 2005; Kalenscher et al., 2005; Moore, Schettler, Killiany, Rosene, & Moss, 2012; Seger & Miller, 2010). However, this heavy emphasis on the role of the cortex in the learning of general rules and the role of the executive in applying these to 'lower-order' processes may be misplaced. Once more, we are faced with difficult questions about the nature of the mechanisms responsible for storing and employing these rules.

This is made more puzzling in the face of evidence that domain-general processing is not restricted to discrete areas of the brain, or even the cortex, but rather that it is distributed across many interacting neural systems (e.g. Ashby & O'Brien, 2005; Poldrack & Foerde, 2008; Seger & Miller, 2010; Smith & Grossman, 2008). Indeed, an extraordinary number of non-primate animals have satisfied experimental behaviour criteria intended to demonstrate the ability to learn and apply even the most complex rules (Katz et al., 2007; Lazareva & Wasserman, 2008; Maes et al., 2015; Minors, 2016; Roitblat & von Fersen, 1992), down to the level of the invertebrate, which have quite substantial differences in neural architecture from the typical subjects of these experiments (Avarguès-Weber & Giurfa, 2013; Minors, 2016).

If one were inclined to invoke the executive, therefore, it should only be a means of last resort. That even very simple animals can demonstrate 'complex' domain-generalisation indicates that the ability can emerge from the specialisation of more phylogenetically basic

neural structures. Indeed, the evolution of the brain is highly conserved. The major features of vertebrate neural organisation have been present from extremely early in its phylogeny (Butler & Hodos, 2005; Holland & Holland, 1999; Katz & Harris-Warrick, 1999; Shu, Hasenstaub, & McCormick, 2003), with homologous properties found in the invertebrate brain (Barron et al., 2015). It would not only be more plausible, but more parsimonious to conclude that this ability is a product of neural attributes that are shared by these creatures, rather than relying on specific neural structures present only in the brains of some animals. As such, exploring how domain-generalisation might emerge from more basic properties of neural function would likely prove a productive enterprise.

3.2.2 Generalisation involves learning the structure of the environment. To start, we should consider again the defining characteristics of these domain-general representations. Ultimately, these 'rules' can be reduced to one shared trait; they comprise some aspect of the statistical structure of the environment. Certain properties of the world (e.g. the compositionality of environmental objects, the functional outcome of behaviours, or the properties of the environment which govern object manipulation [i.e. physics]) drive the development of more domain-general representations when learning these presents a more adaptive solution than relying on highly context sensitive representations (Lake et al., 2016). The crucial characteristic of learning domain-general representations then, is the interaction between environmental structure and the neural systems which process it (Reber, Gitelman, Parrish, & Mesulam, 2003; Seger & Miller, 2010; Zeithamova, Maddox, & Schnyer, 2008). As a consequence, the formation of 'rules' in the sense described above could be present in

numerous neural systems, to the extent that they are responsible for interfacing with the environment. [17]

Let us then return to the neural network literature, which has succeeded in emulating quite complex categorical processes, deriving structure from only the properties of the learning environment.

3.3 Neural Network Architectures Provide Neurally Plausible Mechanisms for Learning the Statistical Structure of the Environment

Proponents of neural network models typically examine how the structure of networks may influence the structure of information processing. As such, rather than characterise task rules and concepts as explicit knowledge structures, they tend to view them as epiphenomenal: emergent products of network dynamics in processing environmental features (McClelland et al., 2010). Recall that in a neural network model, representations take the form of connection weights, which describe what outputs are possible in the context of specific input activation patterns. In biological terms, when environmental contexts activate these representations, certain responses are intrinsically afforded within the network.

3.3.1 Supervised network learning naturalistically supports generalisation across contexts. When neural network models are applied to (highly stylised) cognitive phenomena, they are typically implemented in a supervised manner (LeCun et al., 2015). This means that a generic network architecture will be applied to a task and the experimenter will add information to assist the system in its learning. The connection weights will then adjust in a mimicry of Hebbian plasticity so that the neural network representations will eventually

[17] This may not be true in all cases. For example, some rule-following appears to be linguistically mediated (e.g. Cole, Bassett, Power, Braver, & Petersen, 2014). How the application of a truly external rule to internal processes can be explained in the context of neural network function is not yet apparent, if possible at all, although interesting inroads are discussed in Schiffer, Siletti, Waszak, and Yeung (2017; see also section 3.7).

specify the correct output without the assistance of the experimenter. To illustrate, a network might be shown a series of images of vehicles, labelled by vehicle type. The system will configure itself such that it can eventually distinguish between these vehicle types, absent the labels. This approach has demonstrated that these neural network architectures can produce something akin to categorical learning. Stimuli which are in some way related activate similar or the same representational apparatus, which can thus support generalisation across stimulus classes, in the form of similarity-based inference (e.g. Forbus, Gentner, & Law, 1995; Hinton et al., 1986).[18]

This has led to proposals that the same occurs in a biological context: the neural structures involved in sensorimotor function could conceivably support more domain-general functioning, assuming a network architecture (e.g. Barsalou, 2008). Neuroimaging provides preliminary support for this, revealing perceptual processing regions of the brain appear to also assume responsibility for more domain-general processing.[19]

3.3.2 Network learning can inform and be informed by 'expectation'. Not all domain-generalisation is informed solely by perceptual input. Rather, organisms learn a broad range of expectations that influence processing (as discussed in section 2.4). Consider, as an example, the mis-step that commonly occurs when stepping on a broken escalator, or when one erroneously anticipates a change in elevation; one's expectation informs one's interpretation of the lay of the land. In a similar way, and particularly evident in the visual literature, it appears that the brain anticipates possible interpretations of ambiguous stimuli

[18] One class of neural network architecture which has proved particularly effective in this regard is known as a convolutional neural network or convnet (LeCun et al., 2015), which will be discussed in more detail at section 3.4.1.

[19] For example, motor regions involved in action execution have been found also responsible for planning and interpretation (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese & Goldman, 1998; Goebel, Khorram-Sefat, Muckli, Hacker, & Singer, 1998; Jeannerod, 1994). Visual regions have been found to be responsible for both perception and imagery (Kosslyn, Thompson, & Alpert, 1997).

based on contextually derived expectations, sampling probabilistically from a statistical distribution of possible percepts (e.g. Hoyer & Hyvärinen, 2003; Moreno-Bote, Knill, & Pouget, 2011; Said & Heeger, 2013; Wilson, Krupa, & Wilkinson, 2000). This process undoubtedly enhances an organism's ability to generalise across contexts, interpreting novel circumstance through the lens of past experience (Gregory, 2005).

A variation on the supervised models outlined in the previous section demonstrates how anticipation might be achieved as a function of network dynamics. Here, an expectation is programmed into the network in the form of a conditional probability distribution. When presented with ambiguous stimuli, the network will invert this distribution model using Bayes' rule, in order to compute the posterior probability (e.g. Dayan, Hinton, Neal, & Zemel, 1995; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Yuille & Kersten, 2006). Utilising top-down connections, it will specify desired states for the associative units in lower levels of the processing hierarchy, attempting to create the patterns it anticipates will be presented to it as input (Clark, 2013; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Hinton, 2007). When complemented by the architectural properties of traditional feed-forward models, these probabilistic distributions can be updated over time to improve their accuracy. Such models have made good progress in deriving the structure of higher-order representations based only on the statistical commonalities present in the environment (see Ghahramani, 2015).

3.3.3 In a network, domain-general representations emerge and are updated using the statistical structure of the environment. Both of these approaches emphasise the likelihood that the brain may be structured in order to take advantage of the statistical commonalities shared across ecological contexts, without consolidating this information in some explicit manner. Indeed, they are complementary: with the addition of a prior

probability distribution, neural networks can optimise their behaviour by learning to match probabilistic outputs to the statistical structure of the training environment (Hinton & Salakhutdinov, 2006; Movellan & McClelland, 1993).

It seems prudent to conclude, therefore, that given a network architecture, the capacity to generalise can emerge organically from the statistical structure of the learning environment in the absence of pre-specified knowledge structures (Hinton & Salakhutdinov, 2006; LeCun et al., 2015; Rogers & McClelland, 2004). We are left to explain how these domain-general representations achieve control in circumstances where they conflict. In the following sections, I will argue that the neocortex provides substantial insight into this question. Let us then briefly turn our attention to the architecture of the neocortex, to explore the extent to which it lends itself to an application of the models we have discussed.

3.3.4 Neocortical architecture can plausibly support domain-general network function. The brain is commonly presumed to operate in a feedforward manner. Hebbian theory of synaptic plasticity assumes that changes in neural architecture are induced by the persistent or repeated stimulation of the postsynaptic cell by the presynaptic cell (Hebb, 1949, p. 62), or the lack thereof (Bienenstock et al., 1982). Neural network models, inspired by this property of brain function, similarly employ feedforward architectures. However, the neocortex is distinct in that each feedforward connection is paired with a feedback connection, the function of which is currently unknown (Gilbert & Li, 2013). While it is presumed that such connections facilitate the top-down modulation of earlier information processing, the precise mechanism is as yet unclear.

In a recent paper, David Heeger (2017) utilises these feedback connections to model a process by which the properties of neural network function outlined in the previous section can be plausibly consolidated in a neocortical circuit. Into a generic neural network

architecture, Heeger adds a feedback connection for every feedforward connection. Each unit is therefore altered locally,[20] a feature which supports the generation of more domain-general representations. While neurons lower in the hierarchy of processing will adjust their responses rapidly in accordance with the perceptual input, those neurons higher in the hierarchy are likely to be both slower and more abstract, having been subjected to increasingly abstract input from lower levels.[21] This then permits a neural network architecture to develop domain-general representations in an emergent fashion; the slower, higher levels will be more broadly responsive to the commonalities shared across the details conveyed by the lower, faster levels.

A final input into each unit, the prior, is designed to simulate prior expectation and memory, and depends on the weighted sum of previous responses over time.[22] This permits the system to propagate an expectation, or otherwise stored representation down the processing hierarchy via the feedback connections.[23] This, then, is functionally equivalent to the influence of a probabilistic distribution on the network,[24] supporting anticipation based on contextually derived expectations.

Of course, the specifics of Heeger's model are open to debate, and the theoretical aspects may prove inaccurate in any number of ways. However, the model serves to

---

[20] This is non-trivial. Local adjustment differs from the common approach to weight adjustment employed by neural network models, known as backpropagation; a non-local process which has been criticised as neurally implausible (see Bengio, Lee, Bornschein, Mesnard, & Lin, 2015).

[21] This particular feature mirrors a hypothesis within the neurological literature. It is thought that fast plasticity in the form of large synaptic weight changes, in neural systems lower in the processing hierarchy train higher neural systems which are slower, with smaller weight changes (Gilbert & Li, 2013; Seger & Miller, 2010).

[22] This dependency allows these expectations to take the form of a neural network representation, which are currently active at the higher, more abstract layers of the model, or might be stored for example, in the hippocampus or other neural regions akin to contemporary accounts of Semon's (1921) 'engram' (see also section 3.7).

[23] The dominance of the prior is determined by state parameter variables explicitly linked to empirical findings concerning the influences of other neural processes thought to impact the system, such as attention, neuromodulation, and neural response variability for example.

[24] Indeed, Heeger explicitly makes this link, demonstrating that the prior input can plausibly be interpreted in terms of a prior probability distribution.

demonstrate that the neocortex has a unique capacity to plausibly achieve whatever form of neural network architecture that is required of it to support the naturalistic development of domain-general representations.

This demonstration is a useful prologue to an examination of another unique feature of the neocortex; that it is optimised according to the structure of naturalistic task demands. This property does much to emphasise the credibility of interpreting cortical function from a neural network perspective.

3.4 The Neocortex is Best Viewed as a Multidimensional Task Map

The neocortex is typically viewed as a two-dimensional sheet of functionally similar vertical cell columns (Fig. 5; Hubel & Wiesel, 1959; Mountcastle, 1997; although c.f. Horton & Adams, 2005). Discrete regions of the neocortex can be distinguished by clusters of cell column activation that appear to emphasise different information domains (Cowey, 1979; Penfield & Rasmussen, 1950; Woolsey, 1952). This organisational feature appears to adhere to a principle of 'like-attracts-like'; regions appear to maximise local functional similarity (Kaas & Catania, 2002; Kohonen, 1982; Rosa & Tweedale, 2005; Saarinen & Kohonen, 1985). 'Like-attracts-like' is most well described in the visual cortex, in which adjacent or overlapping areas of the visual field activate adjacent radial cortical columns (Rosa, 2002). The same can be said of the organisation of the primary motor cortex, long observed to be differentially responsive to discrete body parts (e.g. Jackson, 1873; Schott, 1993). The mechanisms which contribute to this broad characteristic are currently unknown, although current views tend to draw from the mechanisms which drive functionally similar sub-cortical regions (Martinetz, 1993; Rosa, 2002; Rosa & Schmid, 1995). In particular,

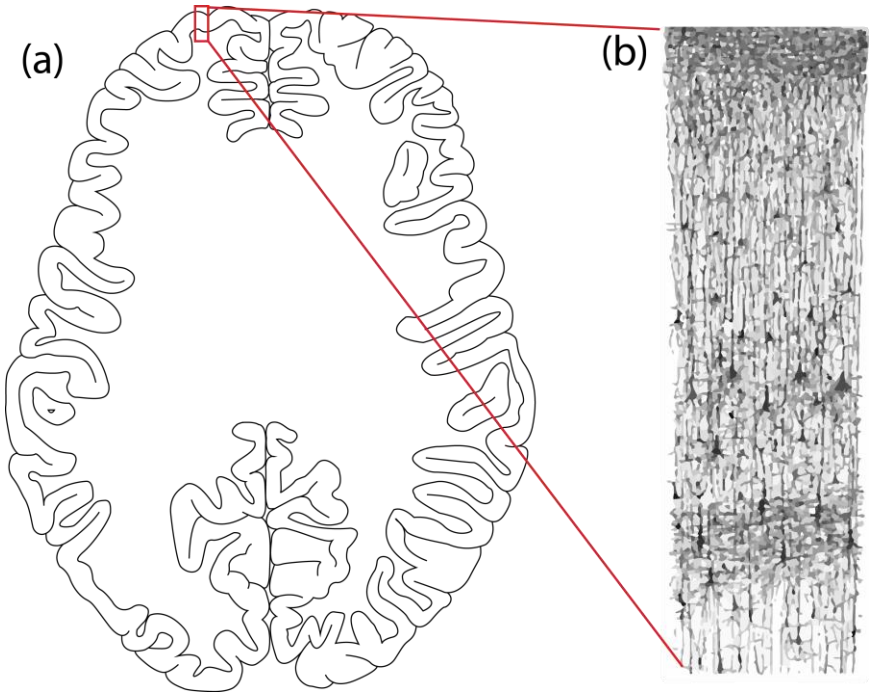

Figure 5. Illustration of neocortical sheet and cortical column. To the left (a), the neocortical sheet of a human brain is exposed along the horizontal plane. The neocortex is comprised of functionally similar vertical cell columns which each span the total height (top to bottom) of the sheet. To the right (b), a segment of the cortical sheet is blown up to illustrate a typical cortical column.

genetically-determined cell-surface chemical cues that guide axons to their targets would naturally result in a localised structure, and activity-dependent synaptic plasticity would result in similar response patterns shared by cells in close proximity. Evolutionary perspectives have also been advocated. For example, a proximally organised architecture would be particularly adaptive due to the increased efficiency borne of the attendant reduction in wiring between associated neurons (Chklovskii, Schikorski, & Stevens, 2002) or the frugal expenditure of action potentials (Barlow, 1961).

Consider again the potential for the neocortex to exploit the properties of neural network function (section 3.3.4). We would expect therefore that the neocortex would also incorporate the structure of the environment to facilitate performance. A recent account of this 'like-attracts-like' characteristic presents a parsimonious interpretation of neocortical organisation that both conforms to our expectations in this regard and encompasses the

considerations outlined above. This view proposes that the topography of the neocortex is best viewed as an optimal mapping of task-relevant parameters onto the two-dimensional (2-D) space of the cortical sheet, in order to maximise interactions between neurons concerned with related operations (Cowey, 1979; Durbin & Mitchison, 1990; Graziano, 2016; Obermayer, Schulten, & Blasdel, 1991).

3.4.1 Mapping visual task parameters to a 2-D plane resembles the visual cortex. In their seminal paper on the topic, Durbin and Mitchison (1990) modelled the connections between neurons restricted to a two-dimensional plane (to mimic the two-dimensional nature of the neocortex) responsive to both stimulus position on the retina and orientation. The model attempted to maximise the colocation of neurons with similar receptive fields. The premise for this constraint was based on a feature of the primary visual cortex (V1) known as 'selective tuning' or 'feature-extraction', in which V1 neurons selectively respond to visual stimulus features (Blakemore, 1974; Niell & Stryker, 2008; Tsotsos et al., 1995). This property is heavily dependent on the interaction between neurons which respond to neighbouring regions of space, an operation which would be prohibitively complex without the colocation of such cells (Cowey, 1979). The result of this 'dimensionality reduction' from the multi-dimensional parameter space to the two-dimensional plane generated a complex arrangement qualitatively resembling that of an experimentally measured macaque striate cortex.

Erwin, Obermayer, and Schulten (1995) extended the model and included binocularity as a parameter, successfully reproducing the relationship between ocular dominance stripes and orientation columns seen in the macaque brain. In parallel, and utilising an alternative dimensionality reduction algorithm, Obermayer and colleagues modelled the same parameters, reproducing experimentally obtained data on cortical ocular

dominance maps and orientation with an extraordinary degree of precision (Obermayer, Blasdel, & Schulten, 1992; Obermayer, Ritter, & Schulten, 1992; Obermayer et al., 1991). Finally, Goodhill and Willshaw (1990) modelled the effect of abnormal eye function in the form of strabismus and monocular deprivation. The model demonstrated that the development of retinotopic stripes characteristic of the visual cortex in primates and the optic tectum in amphibia and fish could plausibly be a function of retinal parameters.

The suggestion then, is that these models do indeed reproduce the ontological pressures which guide neocortical development. The propensity of the neocortex to maximise the colocation of neurons with similar receptive fields results in the mapping of task-relevant parameters to the more constrained 2-D cortical sheet. Thus, an animal's interaction with the environment appears to drive cortical localisation in a manner consistent with the properties of neural network function outlined earlier (section 3.3.1).

Indeed, based on these properties of visual processing, a class of neural network architecture known as 'convolutional' neural networks organise associative layers into 'feature maps', a function which maps data vectors or parameters from feature space to the two-dimensional surface of the feature map (LeCun et al., 2015). These maps can learn specific stimulus features, and then detect those features across the entirety of the map, independent of a given stimulus' position in feature space (for example, its location on an image, or its position in time). Such networks not only match human performance in a range of categorical tasks (Lake, Zaremba, Fergus, & Gureckis, 2015; Mnih et al., 2015; Peterson, Abbott, & Griffiths, 2016), but also predict primate cortical activation patterns in a manner similar to the dimensionality reduction models described here (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2014; Yamins et al., 2014).

3.4.2 Mapping behavioural parameters to a 2-D plane resembles motor cortices. The work of Graziano and Aflalo (2007; also see Aflalo & Graziano, 2006a, 2006b, 2007) successfully extended the dimensionality reduction procedure to the motor cortex. A series of experiments (Graziano, Aflalo, & Cooke, 2005; Graziano, Cooke, Taylor, & Moore, 2004; Graziano, Taylor, & Moore, 2002) applied electrostimulation to macaque cortical regions over a timescale approximating that of common macaque actions. A number of complex movements were provoked, resembling those found in the animal's normal behavioural repertoire, such as reaching and grasping, or hand to mouth actions. These findings spurred the hypothesis that the parameters relevant to the organisation of the motor cortex are related to: (1) the location of muscle groups on the body, (2) spatial correlates to common movements, and (3) the behavioural commonalities shared between aspects of the movement repertoire (Aflalo & Graziano, 2006b; Graziano & Aflalo, 2007). Implementing a dimensionality reduction model which optimised local continuity between these three competing parameters produced a topography bearing a striking similarity to that of the actual macaque motor cortex. This similarity extended well beyond the primary motor

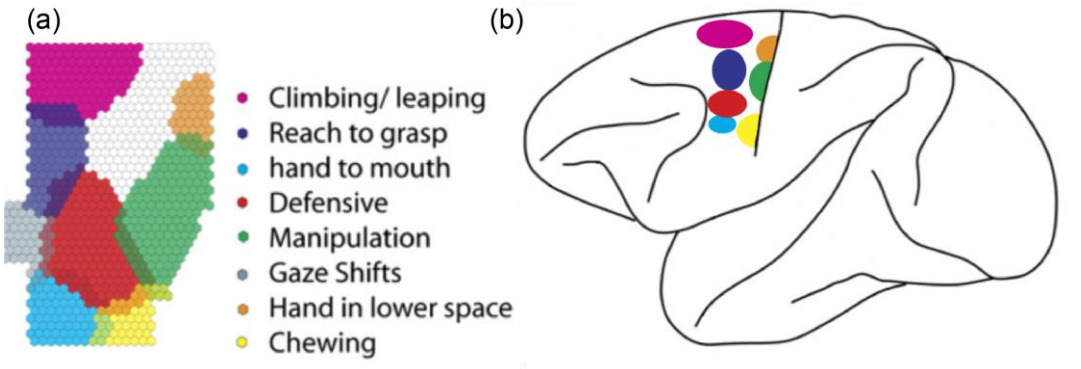

Figure 6. Comparison between model motor map and experimentally obtained data. (a) Arrangement of eight ethological movement categories after model simulation. (b) Activity measured in monkey motor cortex during eight ethological movement categories clustered in these discrete regions. There is a high level of topological similarity between the clustering of the movement categories in both model and actual cortical location. Adapted from Neuron, Vol. 56, Graziano and Aflalo, Mapping Behavioural Repertoire onto the Cortex, pp. 239-251, Copyright (2017), with permission from Elsevier.

cortex, into pre-motor and supplementary regions, reproducing the organisation of roughly 20% of the cortical mantle (Fig. 6).

The results were thus taken to demonstrate that the macaque motor cortex is organised in order to maximise connections between neurons sharing motor control responsibilities, constrained by task demand. The success of this approach contributed to a large body of literature extending these findings across a range of primate and rodent species, incorporating a variety of action-relevant parameters, and including a range of exploratory techniques (see Graziano, 2016 for a review). In all cases, the evidence appears to converge on the motor cortex being organised into zones emphasising common task-related behaviours, constrained by the need to map the multi-dimensional task space to the 2-D cortical sheet.

This 'like-attracts-like' organisational characteristic, then, bears some interesting similarities to the properties of network function described in the previous sections. In co-locating functionally similar neurons, the cortex incorporates the structure of the environment and the interactions of the agent within it. This provides ample opportunity for generalisation in the form of similarity-based inference, assuming a network architecture. The significance of this as it relates to control will be elaborated further in section 3.5, but first I think it appropriate to further highlight the value in interpreting cortical organisation from this perspective as opposed to other, more traditional perspectives.

3.4.3 Mapping task parameters onto the limited dimensions of the neocortical sheet explains neglected features of neocortical topology. While the research in this domain has been restricted to the motor and visual cortex, this theory is uniquely attractive. The 'like-attracts-like' property of the neocortex appears to be operating simultaneously across multiple dimensions of functional similarity and all possible task domains, the

outcome of which is best viewed as a complex multi-dimensional task map. Not only are the empirical results of the dimension-reduction experiments enticing, but the output aligns well with existing findings on the subject, some of which currently remain unexplained.

Importantly, the theoretical underpinnings align with the predominant views regarding the biological mechanisms of cortical development mentioned earlier. Recall that the primary mechanisms thought to drive cortical organisation relate to genetically-determined chemical cues and activity-dependent plasticity (section 3.4). In these models, the localisation function is essentially derived from principles of synaptic plasticity and there is no reason why chemical cues could not contribute. Indeed, noting that the dimension reduction algorithm changes from study to study (at times substantially, e.g. Durbin & Mitchison, 1990; c.f. Obermayer et al., 1991), it seems reasonable to conclude that regardless of the algorithm used, constraining a dimensionality reduction model to maximise local continuity will reproduce actual cortical regions with a high level of similarity. This suggests that any biological mechanism utilising this principle would produce similar results. Likewise, assuming a network architecture, we would expect the neocortex to derive its structure from the properties of the learning environment (section 3.3.4). That the 'like-attracts-like' organisational property naturalistically incorporates the nature of the task-demand is encouraging in this respect.

The resultant output also naturally reproduces features of cortical organisation left largely unaddressed by existing theories. Many researchers note with puzzlement that cortical columns will frequently share overlapping responsibilities, despite belonging to functionally distinct regions of the neocortex (e.g. Graziano, 2008; Rosa & Tweedale, 2005; Schieber, 2001). This is better explained as the result of competitive responsivity to multiple task parameters. Similarly perplexing are the apparently random fractures evident across cortical

regions in which strips of adjacent neurons respond to discrete inputs (e.g. in the motor cortex Manger, Woods, Muñoz, & Jones, 1997; in the visual cortex Rosa & Schmid, 1995) These too can be better explained as an emergent feature of mapping multiple competing task demands onto the two-dimensional surface of the cortex. Both overlapped and fractured organisations naturalistically appear in these models (Aflalo & Graziano, 2006b; Durbin & Mitchison, 1990; Graziano & Aflalo, 2007; Obermayer et al., 1991). Assuming these features are indeed best explained by the colocation of functionally related neurons, they are of critical interest to us in considering the nature of control (section 3.5.1).

Finally, progressive cortical regions appear to respond to stimuli with increasingly complex activation patterns. Traditional approaches to cortical maps interpret these as 'second-order' transformations of stimuli (Allman & Kaas, 1974; Rosa, 2002). Alternatively, they have been considered as a hierarchy in which higher-order areas combine input from lower-order regions responsible for encoding perceptual features of a stimulus, or coordinate the regions responsible for behavioural outputs (Evarts, Shinoda, & Wise, 1984; Porter, 1985). While it is likely that a hierarchical structure does exist to some extent, it is similarly likely that some regions thought to be hierarchical may be better viewed as responding to features of a task which are more complex in nature. For example, the results of the motor cortex models demonstrate that some premotor regions, thought to be hierarchically distinct from the primary motor cortex in primates, may in fact be responsible for emphasising aspects of the animal's movement repertoire requiring more than one body part (Graziano, 2008; Graziano & Aflalo, 2007). Thus, premotor areas necessarily appear more complex than the primary motor cortex, which correlates more closely to individual body parts. Given that both regions project directly to the spinal cord (Dum & Strick, 2002), such an interpretation seems appropriate and indeed, aligns more closely to the rodent literature, which makes no hierarchical distinction between motor areas (see for example Brown & Teskey, 2014;

Dombeck, Graziano, & Tank, 2009; Harrison, Ayling, & Murphy, 2012; Isogai et al., 2012; Ramanathan, Conner, & Tuszynski, 2006). I suggest that this particular characteristic also has implications for the nature of control-related processing, assuming a network architecture (section 3.5.1).

This 'like-attracts-like' characteristic then, presents a particularly attractive lens through which to interpret neocortical organisation. Not only does it align well with existing theory in this domain, it appears to account for previously neglected features of neocortical topography. Crucially, it conforms to our expectations regarding the influence of environmental structure on neural network function; the neocortex derives its structure from the properties of the environment. I suggest that this architectural feature has profound implications regarding the nature of control-related processing.

3.5 Control Emerges from Massively Multiplexed Domain-General Representations.

Assuming a network architecture, a propensity to co-locate neural resources in the manner described above would introduce a substantial level of potential conflict into the system. Recall the notion of *cross talk* (section 3.1). Where two tasks share local resources, they must compete for their use. In the context of the Stroop task, this manifests as a conflict wherein the processes responsible for colour naming and word reading share systems. One must overcome this cross talk to complete the task (Allport, 1955; Cohen et al., 1990). Recall also that neural network models apply this at the level of the neuron in the form of *multiplexing*; where tasks recruit the same representational apparatus to execute multiple tasks, their simultaneous performance becomes impossible (Fig. 4; section 3.1; Feng et al., 2014). In maximising local functional similarity, the likelihood that neocortical regions

will be multiplexed is amplified. The substantial overlap and blurring of functional responsibility characteristic of neocortical maps suggests that this is the case (section 3.4.3).

This is not necessarily surprising. A multiplexed neural architecture would provide substantial flexibility to the system (Botvinick & Cohen, 2014; Feng et al., 2014; Forbus et al., 1995; Hinton et al., 1986). Firstly, multiplexing would minimise the metabolic and functional costs of operation (section 3.4; Barlow, 1961; Chklovskii et al., 2002). More importantly, multiplexing supports generalisation across tasks sharing representations. Recollect that in a neural network, task demands which are in some way related will activate the same representational apparatus, allowing for similarity-based inference (section 3.3.1). This feature is key to the success of neural network models in emulating human performance for complex tasks (section 3.4.1). Multiplexed representations are thus an efficient mechanism with which to generalise across common environmental characteristics to support performance in complex tasks (e.g. Forbus et al., 1995; Hinton et al., 1986). I suggest that the neocortex makes full use of these benefits, particularly as it becomes responsive to increasingly complex task demands.

3.5.1 The neocortex is massively multiplexed. In any system which seeks to minimise redundancy, multiplexing implies an important corollary. As tasks become more complex, they will increasingly share resources with other tasks in order to minimise the double-handling of common sub-processes (Navon & Gopher, 1979; Allport, 1980; Allport, Antonis, & Reynolds, 1972; Logan, 1985; Wickens, 1984). As such, multiplexing would be particularly evident at the level of more abstract, or domain-general representations for at least two reasons.

Firstly, it seems likely that the 'like-attracts-like' principle would manifest at this level. It is no stretch to imagine these as 'higher-level' feature dimensions which would also act as

parameters guiding the topology of neocortical regions. This would complement the finding that some regions of the cortex appear to respond to more complex features of tasks (noted in section 3.4.2; also see Graziano, 2008).

Secondly, recall the implications of neural network learning, particularly as it would pertain to the neocortex (section 3.3.4). As perceptual signals propagate through the system, the neural responses in higher layers will become both slower and more abstract due to the increasing complexity of signals from layers lower in the processing hierarchy (a functionality which compliments neurocognitive proposals, e.g. Seger & Miller, 2010). Consequently, higher levels will be broadly responsive to environmental commonalities, rather than transient features. This means that 'like-attracts-like' is not only influencing higher levels directly, but also via the signals propagated from lower levels, themselves organised according to 'like-attracts-like'.

Thus, as the neocortex becomes responsive to increasingly complex task parameters, it will also become increasingly multiplexed. Given that representational multiplexing will introduce additional conflict in any ecological scenario in which functionally similar tasks, or their supporting internal processes compete, this seems to present a troubling control problem. However, recent work by Feng et al. (2014) suggest it actually limits the complexity required of a control mechanism.

3.5.2 Multiplexing intrinsically constrains control. Feng et al. (2014) modelled a number of different sized neural networks, containing between 10 to 1000 individual representations. Each representation was given an individual 'task demand' unit to implement control by biasing the response of the representation, allowing it to 'win out' over competitors. The system had no intrinsic constraint on control, meaning every control unit could be activated at once. A 'task' in the system was simulated when the output unit activity

of a given representation was congruent with the input unit activation. The authors then assigned each representation a random amount of overlap with other representations in the system to simulate multiplexing.

Two findings are particularly pertinent to this thesis. Unsurprisingly, multiplexing imposed limitations on the multi-tasking of those 'tasks' subject to interference. Interestingly, this limitation reached a maximum number of conflicting tasks after only a modest degree of multiplexing, and upon which the size of the network had only a marginal influence. This suggests that introducing multiplexing into a neural architecture intrinsically limits the amount of possible multitasking quite significantly, regardless of the representational capacity of the system. This implies that multiplexing also limits the complexity required of the control system—it need only provide control in circumstances where multiple tasks are possible.

The second finding was somewhat counter-intuitive. While activating control units did improve performance to a point, it also contributed cross talk to the system (Fig. 7). Connections in the model could be either excitatory or inhibitory. Multiplexing resulted in the same output units containing both inhibitory and excitatory links from different input units. Recall that 'tasks' required the output units of a representation to be congruent with the input units. Should a control unit increase the strength of an irrelevant representation possessing an incongruent connection to the output units involved in the task, this would interfere with the congruency between the input and output units of the task. For example, in a task requiring a representation with excitatory links from input to output units, allowing a control unit to strengthen an irrelevant representation with an inhibitory link to the output units involved in the task would interfere with the task's performance. As such, to achieve

optimal performance, the system was forced to limit the number of active control units, else introduce too much additional conflict into the system.

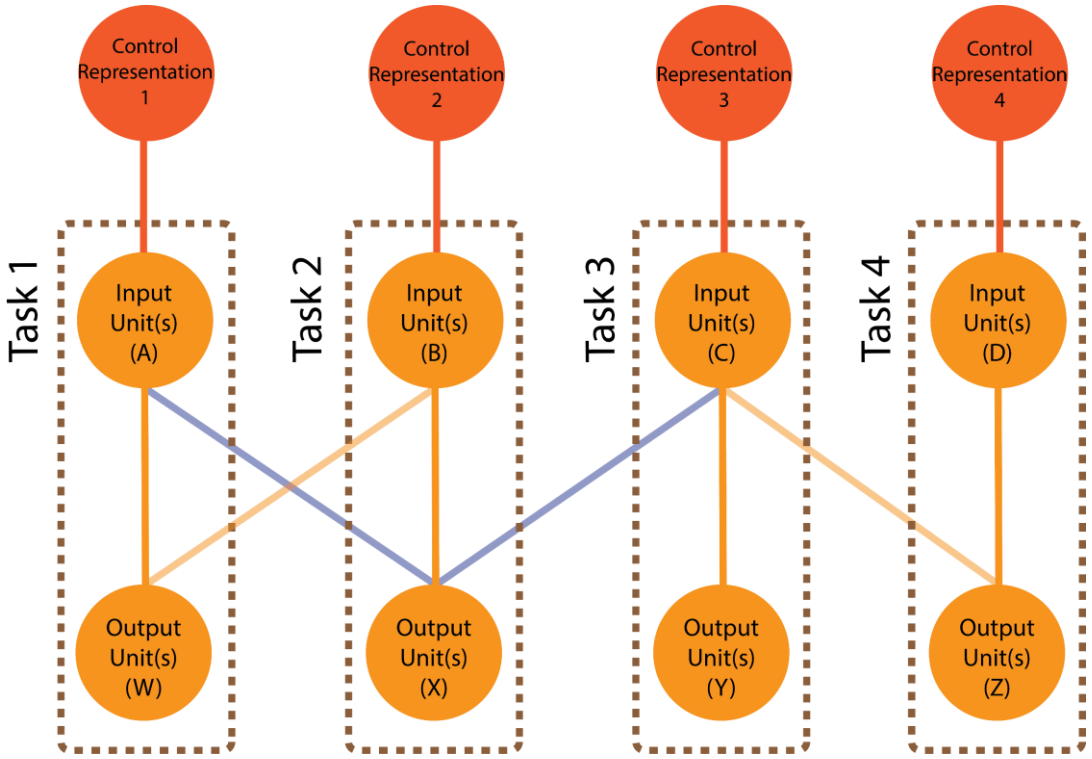

Figure 7. Control units may contribute cross talk by activating irrelevant multiplexed representations with incongruent connections to a task. This diagram illustrates four tasks (brown dashed lines) which engage multiplexed representations (diagonal connection represents multiplexing). Orange connections are excitatory. Blue connections are inhibitory. Activating Control Representations 3 and 4 to increase the responding of C and D would have no effect on the ability to carry out Tasks 3 and 4 simultaneously as all connection types are congruent. However, the activation of Control Representation 3 would introduce conflict into Task 2 as the connection between C and X is incongruent with the connection between the input units (B) and output units (X) involved in the task.

When considered together, the findings indicate that not only would control in a network architecture prove deleterious in some respects, the presence of multiplexing alone would intrinsically limit the number of control-demanding tasks that could be carried out simultaneously. Thus, the control system need only provide control in very limited circumstances: where multiple tasks are possible. This then begs the question, in a neural network architecture constrained by the implications of multiplexing, under what conditions *would* the system need to allocate control?

3.5.3 The need for control is determined by task pre-conditions. The final piece of the puzzle concerns a defining characteristic of complex behaviour. Recall that cognitive control describes the tuning of cognitive processes in service of a goal. It has been the recurring observation of behavioural scientists that purposive behaviour takes the form of a pre-conditional hierarchy: simple actions are coordinated to achieve subtasks, which themselves are nested within more complex tasks (Lashley, 1951; Miller, Galanter, & Pribram, 1986; Sacerdoti, 1974). As a simplistic example, consider the hungry monkey. To reach the banana, it must first climb the tree. To climb the tree, it must coordinate reaching and grasping movements of the hands and feet. Pre-conditional hierarchies are a strikingly ubiquitous feature of goal-directed behaviour; even the humble bumblebee has been observed to manipulate objects in order to reach a sugar water reward (Alem et al., 2016). More to the point, experimental evidence suggests that complex animals deconstruct these hierarchies to inform their planning (Bruner, 1973; Fischer, 1980; Greenfield & Schneider, 1977; Zacks, Kurby, Eisenberg, & Haroutunian, 2011).

This should not surprise us. The benefits to an animal in exploiting the extant pre-conditional hierarchies in naturalistic tasks are identical to the benefits of grouping stimuli and responses into functional categories. By decomposing complex behaviours into constituent parts, not only can they be more efficiently coded, but they also lend themselves to generalisation—the constituents can be applied in different ways under different conditions in ways that a larger behavioural routine cannot (Allport & Alan, 1997; Hayes-Roth & Hayes-Roth, 1979; Laird, Rosenbloom, & Newell, 1986; Taatgen & Lee, 2003). We

should expect then that as a defining parameter of naturalistic tasks, the neocortex will incorporate the pre-conditional nature of behaviour into its task map.[25]

The evaluation of the influence of pre-conditional hierarchies on the physical topology of the brain is in its infancy. However, the research appears to converge on a structural feature that not only conforms to our expectations, but has great significance for the implications brought forward by Feng et al. (2014), as described in the previous section. Pre-conditional hierarchies will characteristically demonstrate what has been called a 'community structure' (Newman, 2004; Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004); groups of related items (a community) will be distinguishable from other communities by the pre-conditions which define them (Botvinick & Cohen, 2014).

An exemplar in this respect is the Tower of Hanoi task (Fig. 8), commonly employed in cognitive control assays (Allport & Alan, 1997). Visualised in graph form the task manifests as a network of densely connected groups, representing the legal moves permissible in each stage of the task, separated by bottle-neck like edges, reflecting the necessary pre-conditions for exploiting the legal moves in each stage and thus, solving the task.

[25] The cortical localisation literature discussed in section 3.4 emphasises genetically-determined cell-surface chemical cues which guide axons to their targets and activity-dependent synaptic plasticity as likely biological contributors to the localised structure apparent in the neocortex. These same mechanisms could be invoked to account for the mapping of pre-conditional structures to the cortex. Accounts of motor behaviour tend to characterise action sequences as originating from genetically specified propensities (e.g. Aldridge & Berridge, 1998; Bruner, 1973; Elfwing, Uchibe, Doya, & Christensen, 2007) and developed further via associative learning, which is facilitated in the brain by Hebbian plasticity (e.g. Conway & Christiansen, 2001; Fischer, 1980; Greenfield, Nelson, & Saltzman, 1972).

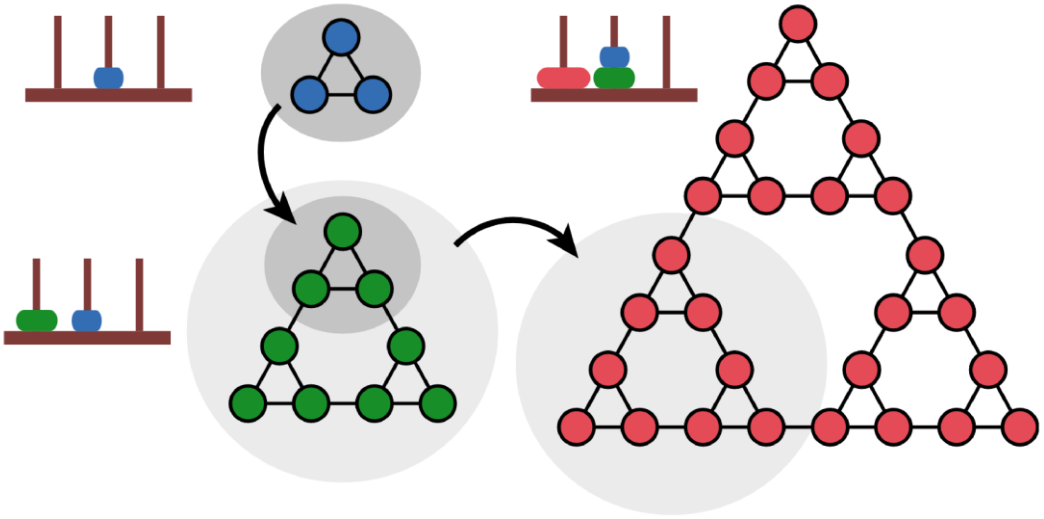

Figure 8. Tower of Hanoi task visualised in graph space. The Tower of Hanoi involves a set of disks of varying sizes, each of which can be placed on any of three posts. No disk can be placed on a smaller disk. With the smallest (blue) disk only, there are three options. When a second, larger (green) disk is added, a pre-condition becomes apparent. Blue can make three moves for every position green takes on the posts, but green can only make one movement (to the post that blue does not occupy). Adding a third, still-larger (red) disk adds a second series of pre-conditions. The blue and green movements are unconstrained by red, but red can only move to the post that green and blue do not occupy. The resulting structure illustrates the general point that pre-conditional tasks naturally induce hierarchical structures, comprising clusters or "communities" of states separated by state space bottlenecks. Reprinted from Cognitive Science, Vol. 38, Botvinick and Cohen, The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers, pp. 1249-1285, Copyright (2017), with permission from John Wiley and Sons.

The brain appears to be highly sensitive to the community structure of such naturalistic tasks. Regions of highly localised activity, which respond to discrete task features (termed *modules*),[26] coactivate differentially during both task and resting states (termed *functional connectivity*). Contemporary accounts of this functionality suggest that the brain is primarily organised to reflect a relatively stable community structure in a resting state, with only minor deviations during some task states (see Cole et al., 2014 for a review). A recent meta-analysis of more than 1600 functional connectivity studies provides strong evidence for this assessment, estimating the similarity of activation patterns across 638 brain regions during experimental tasks (Crossley et al., 2013). The authors found a high correlation

[26] While these 'modules' have at times been equated with Fodor's (1983) mental modules (e.g. Meunier, Lambiotte, & Bullmore, 2010), given the extensive evidence intimating the like-attracts-like properties of the cortex (see again section 3.4, 3.5.1), I suggest that these are better considered in terms of localised regions of functional similarity.

between the resultant data and the resting state neuroimaging data of healthy control subjects. This stability conforms well to our expectations assuming a network derived task map – at higher levels of processing, neural responsivity should be more broadly responsive to environmental commonalities, rather than transient features (section 3.5.1).

Interestingly, the authors identified apparent competition between nodes, signified by concomitant activation and deactivation of paired regions during tasks. These competitive interactions were more likely to appear between regions in different modules than those in the same module. More importantly, this was most apparent in the relationship between 21 particularly high-density *rich-club* nodes which were responsive to a larger set of task features, and less well connected *peripheral* nodes, responsive to a more restricted number of task features. Specifically, activation in rich-club nodes was accompanied by decreases in the activation of peripheral nodes. The authors suggest two alternative explanations for this phenomenon. One interpretation is that coactivation is a finite resource; all modules cannot be simultaneously coactive (see also Kastner, De Weerd, Desimone, & Ungerleider, 1998). The alternative is that rich-club nodes 'switch off' modules under certain task conditions, or suppress activity in inactive modules (see also Drevets et al., 1995; Kawashima, O'Sullivan, & Roland, 1995). While the authors critique the latter as homuncular, I suggest that in the context of network function, these accounts are complementary. Consider the implications of multiplexing (section 3.5.2). The activation of domain-general representations is likely to intrinsically constrain the activation of functionally similar representations. The associated deactivations in peripheral nodes may well come about as a result of representational multiplexing. However, we must also consider the proposed function of domain-general representations—to detail the functional relevance of behaviours across a broader range of environmental features and bias responding accordingly (section 3.1.1; 3.2.2). It may be that

the associated deactivations come about as the more domain-general rich-club nodes bias responding to allocate multiplexed resources to one task over another (see also section 3.5.4).

Recent computational work complements these findings, providing insight into how these pre-conditional task 'modules' are formed and implemented. Applying a generic neural network model to a task with a community structure will reflect that structure following training; nodes within a sub-task grouping will be reflected by similar activation patterns than those in different groups (Schapiro, et al. 2013). Neuroimaging reveals an equivalent effect within the frontal and temporal cortex (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013) and hippocampus (Schapiro, Turk-Browne, Norman, & Botvinick, 2016) of human participants during the same or similar tasks; neural responsivity to events clusters according to sub-task groupings with training. Further, when a model network is imbued with a probabilistic distribution at the highest level of processing (as seen in section 3.3.3), it will demonstrate simulated behaviour very similar to that demonstrated by humans in real behavioural assays (Collins & Frank, 2013; Mnih et al., 2015; Solway et al., 2014). Both simulated agent and human participant are sensitive to and utilise bottlenecks in their planning, indicating that humans may possess an expectation that tasks can be deconstructed to exploit naturalistic pre-conditions.

3.5.4 A community structure implies that when tasks conflict, control will be emergent and functionally segregating. This sensitivity to community structure then, quite neatly describes the conditions which would demand the allocation of control. In circumstances where task parameters congregate on cortical regions of particular functional similarity, representational multiplexing will inherently prevent these tasks from being carried out simultaneously. These regions would be determined in part by their role as a pre-conditional bottleneck: where multiple tasks require the operation of that cortical zone as a

pre-condition to their execution. Thus, when two or more tasks compete for that resource in a given ecological scenario, the nature of the task, informed by the domain-general representations specifying the appropriate contingency-context pairing, will bias responding to allocate the resource to one task over another. The plausibility of this conclusion is highlighted by the findings of Crossley et al. (2013): highly interconnected modules of functional similarity appear to compete for activation. When more domain-general neocortical regions are active, less well connected peripheral regions deactivate.

Recall the task demand units posited by Cohen et al. (1990); independent domain-general control units which serve to bias processing according to the nature of the task demand (section 3.1.1). Miller and Cohen's (2001) *guided activation theory* (GAT) extrapolates this idea in a manner quite compatible with the conclusion proposed here. In GAT, it is suggested that the prefrontal cortex can be viewed as a map that specifies the pattern of neural network representations in other cortical or subcortical regions required to solve a task, particularly when these pathways overlap. Botvinick, Niv, and Barto (2009) review evidence suggesting the dorsolateral striatum may act in a similar manner, drawing inputs of task sets from frontal cortices and specifying neural network representations accordingly, with particular reference to the region's response specificity during action sequences. Both proposals have seen successful application in the context of a neural network architecture, emphasising the neural plausibility of these accounts (Botvinick & Cohen, 2014; Frank & Claus, 2006; O'Reilly & Frank, 2006; Rougier et al., 2005). Indeed, GAT is but one of a number of recent theories that share a similar premise (see Botvinick & Cohen, 2014 for an overview). [27] Despite their differences in mechanism, the common thread concerns more general, established representations that are activated to the appropriate degree in order to

[27] Although GAT has seen the most substantial empirical development.

bias less complex representations in service of an adaptive outcome. I suggest that in many cases, the pre-conditional representations described in the preceding section not only serve this purpose, but also serve as the signal that control is needed, activated when multiple tasks require them as a resource.

3.6 Conclusion: Control is Achieved as a Function of the Pre-conditions for Representational Action Sets

Let us then return to Botvinick and Cohen's (2014) 'metaoptimization' problem; how might control mechanisms optimise performance across a broader range of tasks most efficaciously? A salient corollary concerns the nature of this optimisation; how can we account for these control mechanisms without deferring to the homunculus? Quite profound answers to this question appear plausible in the neocortex.

Complex behaviour seems to necessitate representations which permit an animal to generalise performance across tasks sharing environmental commonalities. These would be far more adaptive than the proliferation of representations accounting for each individual pairing of context to action set. Assuming a network architecture, neural network function provides substantial insight into the likely nature of these domain-general representations. The statistical structure of the environment, and the nature of the animal's interactions therein, would naturalistically drive the development of complex neural representations which map action sets to the appropriate contextual contingencies. The responses of neurons lower in the processing hierarchy would adjust their responses rapidly in accordance with sensorimotor input, while those neurons higher in the hierarchy are likely to be both slower and more abstract, having been subjected to increasingly abstract input from lower levels. These higher levels would consequently be more responsive to the commonalities

shared across the environmental details conveyed by the lower, faster levels. This would support similarity-based inference as a form of generalisation.

This property appears to be realised in the neocortex. Not only does it possess the architectural features required to develop these representations, it is also characterised by an organisational propensity that emphasises the commonalities between task-related parameters. This 'like-attracts-like' characteristic closely aligns with our expectations assuming a network architecture. However, in co-locating functionally similar neurons, the neocortex will be subject to representational multiplexing: where tasks recruit the same representational apparatus to execute multiple tasks, their simultaneous performance becomes impossible. Indeed, as the neocortex becomes responsive to increasingly complex task parameters, it will become increasingly multiplexed. It is from this constraint that we derive the optimisation criterion for domain-general control-related processing.

In such an architecture, the presence of multiplexing alone would intrinsically limit need for control to only those circumstances where multiple tasks are possible. These circumstances are defined by the pre-conditional structure of naturalistic tasks, a property of the learning environment that is robustly featured in the organisation of the neocortex. Behaviours which serve as the preconditions for multiple action-sets manifest in the brain as densely-connected cortical regions responsive to a broad array of task demands. Where multiple tasks compete for this resource, these domain-general representations specify the pattern of neural network representations in other cortical or subcortical regions according to the nature of the task demand.

On this account of control, there is no need to posit the existence of explicit knowledge structures to explain an animal's ability to discern the functional relevance of behaviours in increasingly complex environments. Nor is there a need to appeal to the

executive mechanisms which feature so prevalently in the neurocognitive literature. Rather, the development of domain-general representations may merely be an epiphenomenal product of the interaction between the statistical structure of the environment and the neural systems which process it. These representations would serve as both the signal that control is required and the mechanism by which conflicts are resolved. Crucially, these representations are entirely plausible in the context of the nervous system.

3.7 Outlook for the Future

Of course, this account merely sets the stage for further enquiry into the possibilities for network function to provide insight into the neural bases of complex behaviour. As such, before I conclude, let us briefly explore the prospects for future research in light of the material discussed in chapter 3. In my mind, in addition to the ongoing research into Guided Activation Theory and its brethren (section 3.5.4) three avenues of empirical pursuit hold particular value.

The cortex appears to be particularly suitable for generating the kinds of representations that can achieve control in more complex tasks (see section 3.3.4). Yet, animals which lack a cortex are not only capable of organising behaviour around functional categories (e.g. Avarguès-Weber & Giurfa, 2013; Giurfa, 2007; Hoy, 1989; Windhorst, 1996), but they do so in a manner that demonstrates the exploitation of pre-conditions (e.g. Alem et al., 2016). The approaches to domain-generalisation in the classical neurocognitive literature tend to characterise rule-learning as a hierarchy of increasingly complex abilities (section 3.2). These abilities are commonly used as a comparative measure of animal intelligence (Katz et al., 2007; Minors, 2016; Zayan & Vauclair, 1998). Exploring the features of neural architecture that could support the functionality discussed in this thesis in non-cortical regions, or their homologues in invertebrates might provide useful insights in this domain.

The next avenue concerns the nature of expectations in facilitating control. Our interpretation of perceptual input is at times influenced by contextually derived expectations; anticipated states, and their dependencies on earlier states and actions. Sections 3.3.2 and 3.3.4 suggest ways in which these might contribute to control-related processing, but relies heavily on the assumption that these are stored and retrieved. No doubt this may be partially accounted for in the context of the pre-conditional representations outlaid in Section 3.5. Yet, expectations also facilitate control mechanisms in a manner that does not presuppose cortical intervention (see section 2.4.2). Indeed, behavioural and electrophysiological data increasingly implies the presence of enduring temporal and spatial representations (e.g. Campos, Cherian, & Segraves, 2006; Foster & Wilson, 2006; Jirenhed, Rasmussen, Johansson, & Hesslow, 2017; Menzel et al., 2005; Moser, Rowland, & Moser, 2015; Pfeiffer & Foster, 2013; Spencer & Ivry, 2013; Yoganarasimha, Yu, & Knierim, 2006). Linking the material discussed in this thesis to the vast literature on memory may profitably advance our knowledge in this domain. I suggest a useful starting point may lie in interpreting contemporary accounts of Semon's (1921) 'engram', through the lens of network dynamics. Engrams, or 'memory traces', are thought to be (semi)permanent neural changes that explain the persistence of memory, typically thought to be widely distributed throughout the brain (Bruce, 2001; Schacter, 1996). While their exact mechanism has proven elusive, I suggest that inroads may exist in the literature on neural consolidation (Dudai, 2004).

One must also eventually face the role of language in mediating control. Much has been made of the human ability to rapidly adjust behaviour in response to instructions (e.g. Cole et al., 2014). This capacity is not limited to humans, or even primates (e.g. Pepperberg, 1987). How the application of what is truly an external rule to internal processes can be explained in the context of neural network function is not yet apparent, if possible at all, although interesting inroads are discussed in Schiffer et al. (2017). Early thoughts on this

posed the notion that the ability to form a mental representation of a concept via language would help to explain this capacity (Premack, 1978; Premack & Premack, 1983; Thompson & Oden, 1995). Attempting to explain these in the context of neural network representations may prove a fruitful exercise.

Conclusion

We commenced this thesis asking the question, how can we understand the phenomena of cognitive control in a manner that is compatible with a neurally plausible perspective on representation? By restricting our perspective to the kinds of information currently explicable in a neural context, myriad opportunities emerge for the brain to resolve conflicts without appealing to the inscrutable executive mechanisms proposed in the classical cognitive literature. Foremost, many of the traditional components thought to comprise control can be achieved by quite basic features of neurally plausible network dynamics. These networks would require only the evolutionarily-hardwired, or experientially-learned valence of stimuli to adaptively resolve a broad array of conflicts and coordinate responding in service of a goal. Of course, as animals become more sophisticated, so too do their mechanisms for achieving control. Yet, the kinds of mechanisms that can resolve conflicts evoked by more complex tasks in a neurally plausible manner appear to again be quite different from their counterparts in the cognitive literature. The neocortex appears to develop quite intricate representations from the structure of an animal's interactions with the environment that serve as both the source of and solution to neural conflicts. To the extent that this interpretation of neural function proves true, it would have profound implications for the nature of complex behaviour that extend far beyond the phenomena of cognitive control.

References

Adler, J. (1966). Chemotaxis in bacteria. *Science, 153*(3737), 708-716. doi:10.1126/science.153.3737.708

Aflalo, T. N., & Graziano, M. S. (2006a). Partial tuning of motor cortex neurons to final posture in a free-moving paradigm. *Proceedings of the National Academy of Sciences of the United States of America, 103*(8), 2909-2914. doi:10.1073/pnas.0511139103

Aflalo, T. N., & Graziano, M. S. (2006b). Possible origins of the complex topographic organization of motor cortex: reduction of a multidimensional space onto a two-dimensional array. *Journal of Neuroscience, 26*(23), 6288-6297. doi:10.1523/jneurosci.0768-06.2006

Aflalo, T. N., & Graziano, M. S. (2007). Relationship between unconstrained arm movements and single-neuron firing in the macaque motor cortex. *Journal of Neuroscience, 27*(11), 2760-2780. doi:10.1523/jneurosci.3147-06.2007

Aldridge, J. W., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *Journal of Neuroscience, 18*(7), 2777-2787.

Alem, S., Perry, C. J., Zhu, X., Loukola, O. J., Ingraham, T., Søvik, E., & Chittka, L. (2016). Associative Mechanisms Allow for Social Learning and Cultural Transmission of String Pulling in an Insect. *PLoS biology, 14*(10), e1002564. doi:10.1371/journal.pbio.1002564

Allman, J. M., & Kaas, J. (1974). The organization of the second visual area (V II) in the owl monkey: a second order transformation of the visual hemifield. *Brain research, 76*(2), 247-265.

Allman, J. M., & Martin, B. (1999). *Evolving brains*. New York: Scientific American Library.

Allport, F. H. (1955). Theories of perception and the concept of structure: A review and critical analysis with an introduction to a dynamic-structural theory of behavior. *Optometry and Vision Science, 33*(4). doi:10.1037/11116-000

Allport, G. W., & Alan. (1997). Planning and problem solving using the five disc Tower of London task. *The Quarterly Journal of Experimental Psychology: Section A, 50*(1), 49-78. doi:10.1080/713755681

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature, 397*(6715), 168-171. doi:10.1038/16483

Anderson, J. R. (2013). *The Architecture of Cognition*. New York: Psychology Press.

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681-692. doi:10.1016/j.neuron.2005.04.026

Anthony, W. (1959). The Tolman and Honzik insight situation. *British Journal of Psychology, 50*(2), 117-124. doi:https://search.proquest.com/docview/1293607565

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-Specific Neural Activity in the Primate Prefrontal Cortex. *Journal of Neurophysiology, 84*(1), 451-459. Retrieved from http://jn.physiology.org/content/84/1/451

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences, 9*(2), 83-89. doi:10.1016/j.tics.2004.12.003

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*, 403-450. doi:10.1146/annurev.neuro.28.061604.135709

Avarguès-Weber, A., & Giurfa, M. (2013). Conceptual learning by miniature brains. *Proceedings of the Royal Society B, 280*(1772), 20131907. doi:10.1098/rspb.2013.1907

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47-89): Academic Press.

Badgaiyan, R. D. (2000). Executive control, willed actions, and nonconscious processing. *Human brain mapping, 9*(1), 38-41. doi:10.1.1.405.5013

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological bulletin, 121*(1), 65.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication* (pp. 217-234). Cambridge: M.I.T. Press.

Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality, 40*(1), 35-55.

Barron, A. B., Gurney, K. N., Meah, L. F. S., Vasilaki, E., & Marshall, J. A. R. (2015). Decision-making and action selection in insects: inspiration from vertebrate-based theories. *Frontiers in Behavioral Neuroscience, 9*(216). doi:10.3389/fnbeh.2015.00216

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*(1), 617-645. doi:10.1146/annurev.psych.59.103006.093639

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences, 107*(50), 21767-21772. doi:10.1073/pnas.0908104107

Bazhenov, M., Huerta, R., & Smith, B. H. (2013). A computational framework for understanding decision making through integration of basic learning rules. *The Journal of Neuroscience, 33*(13), 5686-5697. doi:10.1523/jneurosci.4145-12.2013

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind*. London: Oxford: Blackwell.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience, 2*(1), 32.

Blakemore, C. (1974). Developmental factors in the formation of feature extracting neurons. In *The neurosciences: Third study program* (pp. 105-114). Cambridge: MIT Press.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700. doi:10.1037/0033-295X.113.4.700

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural computation, 19*(2), 442-477. doi:10.1162/neco.2007.19.2.442

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review, 108*(3), 624-652. doi:10.1037//0033-295X.108.3.624

Botvinick, M. M., & Cohen, J. D. (2014). The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers. *Cognitive Science, 38*(6), 1249-1285. doi:10.1111/cogs.12126

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. doi:10.1016/j.cognition.2008.08.011

Brandman, O., & Meyer, T. (2008). Feedback loops shape cellular signals in space and time. *Science, 322*(5900), 390-395. doi:10.1126/science.1160617

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews, 26*(7), 809-817. doi:10.1016/S0149-7634(02)00067-2

Bray, D. (2009). *Wetware: a computer in every living cell*. New Haven: Yale University Press.

Brody, C. D., & Hanks, T. D. (2016). Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology, 37*, 149-157. doi:10.1016/j.conb.2016.01.003

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron, 68*(5), 815-834. doi:10.1016/j.neuron.2010.11.022

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems, 6*(1), 3-15. doi:10.1016/S0921-8890(05)80025-9

Brooks, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge: MIT Press.

Brown, A. R., & Teskey, G. C. (2014). Motor cortex is functionally organized as a set of spatially distinct representations for complex movements. *Journal of Neuroscience, 34*(41), 13574-13585. doi:10.1523/jneurosci.2500-14.2014

Bruce, D. (2001). Fifty years since Lashley's In search of the Engram: refutations and conjectures. *Journal of the History of the Neurosciences, 10*(3), 308-318. doi:10.1076/jhin.10.3.308.9086

Bruner, J. S. (1973). Organization of early skilled action. *Child development*, 1-11.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology, 61*, 27-48. doi:10.1146/annurev.psych.60.110707.163508

Bures, J., Bermúdez-Rattoni, F., & Yamamoto, T. (1998). *Conditioned taste aversion: Memory of a special kind*. London: Oxford University Press.

Butler, A. B., & Hodos, W. (2005). *Comparative vertebrate neuroanatomy: evolution and adaptation*. New Jersey: John Wiley & Sons.

Campos, M., Cherian, A., & Segraves, M. A. (2006). Effects of Eye Position upon Activity of Neurons in Macaque Superior Colliculus. *Journal of Neurophysiology, 95*(1), 505-526. doi:10.1152/jn.00639.2005

Carr, L., Iacoboni, M., Dubeau, M.-C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences, 100*(9), 5497-5502. doi:10.1073/pnas.0935845100

Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature neuroscience, 7*(3). doi:10.1038/nn1194

Carruthers, P. (in press). Valence and Value. *Philosophy and Phenomenological Research*. doi:doi: 10.1111/phpr.12395

Chemero, A. (2000). Anti-Representationalism and the Dynamical Stance. *Philosophy of Science, 67*(4), 625-647. doi:10.1086/392858

Chklovskii, D. B., Schikorski, T., & Stevens, C. F. (2002). Wiring Optimization in Cortical Circuits. *Neuron, 34*(3), 341-347. doi:10.1016/S0896-6273(02)00679-7

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*, 73-101. doi:10.1146/annurev.psych.093008.100427

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge: MIT Press.

Cisek, P. (2001). Embodiment is all in the head. *Behavioral and Brain Sciences, 24*(01), 36-38. doi: 10.1017/S0140525X0124391X

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1485), 1585-1599. doi:10.1098/rstb.2007.2054

Clark, A. (1995). Moving Minds: Situating Content in the Service of Real-Time Success. *Philosophical Perspectives, 9*, 89-104. doi:10.2307/2214213

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. doi:10.1017/S0140525X12000477

Clark, A., & Toribio, J. (1994). Doing without Representing? *Synthese, 101*(3), 401-431.

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A Systems-Level Perspective on Attention and Cognitive Control: Guided Activation, Adaptive Gating, Conflict Monitoring, and Exploitation versus Exploration. In M. I. Posner (Ed.), *Cognitive Neuroscience of Attention* (pp. 71-90). New York: Guilford Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review, 97*(3), 332. doi:10.1037/0033-295X.97.3.332

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron, 83*(1), 238-251. doi:10.1016/j.neuron.2014.05.014

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review, 120*(1), 190. doi:10.1037/a0030852

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex, 10*(9), 910-923. doi:10.1093/cercor/10.9.910

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*(2), 89-97. doi:10.1080/00207727008920220

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences, 5*(12), 539-546. doi:10.1016/S1364-6613(00)01800-3

Cooper, R. P. (2010). Cognitive Control: Componential or Emergent? *Topics in Cognitive Science, 2*(4), 598-613. doi:10.1111/j.1756-8765.2010.01110.x

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114. doi:10.1017/S0140525X01003922

Cowey, A. (1979). Cortical maps and visual perception the grindley memorial lecture. *The Quarterly journal of experimental psychology, 31*(1), 1-17. doi:10.1080/14640747908400703

Craik, K. J. W. (1967). *The nature of explanation* (Vol. 445). Cambridge: Cambridge University Press.

Crossley, N. A., Mechelli, A., Vértes, P. E., Winton-Brown, T. T., Patel, A. X., Ginestet, C. E., . . . Bullmore, E. T. (2013). Cognitive relevance of the community structure of the human

brain functional coactivation network. *Proceedings of the National Academy of Sciences, 110*(28), 11583-11588. doi:10.1073/pnas.1220826110

Damasio, A. R. (1999). *The feeling of what happens: body and emotion in the making of consciousness*. Boston: Mariner Books.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience, 8*(4), 429-453. doi:10.3758/cabn.8.4.429

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation, 7*(5), 889-904. doi:10.1162/neco.1995.7.5.889

Dennett, D. (1987). *The Intentional Stance*. Cambridge: MIT Press.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222. doi:10.1146/annurev-psych-122414-033400

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*(1), 176-180. doi:10.1007/bf00230027

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 308*(1135), 67-78.

Diekamp, B., Kalt, T., & Güntürkün, O. (2002). Working memory neurons in pigeons. *The Journal of Neuroscience, 22*(4), RC210. Retrieved from http://www.jneurosci.org/content/22/4/RC210.full.pdf

Ditterich, J., Mazurek, M. E., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature neuroscience, 6*(8), 891-898. doi:10.1038/nn1094

Dombeck, D. A., Graziano, M. S., & Tank, D. W. (2009). Functional clustering of neurons in motor cortex determined by cellular resolution imaging in awake behaving mice. *Journal of Neuroscience, 29*(44), 13751-13760. doi:10.1523/jneurosci.2985-09.2009

Doya, K. (2008). Modulators of decision making. *Nature neuroscience, 11*(4), 410. doi:10.1038/nn2077

Drevets, W. C., Burton, H., Videen, T. O., Snyder, A. Z., Simpson, J. R., & Raichle, M. E. (1995). Blood flow changes in human somatosensory cortex during anticipated stimulation. *Nature, 373*(6511), 249-252. doi:10.1038/373249a0

Dreyfus, H. L. (2002). Intelligence without representation: the relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1*(4), 367-383.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology, 55*, 51-86. doi:10.1146/annurev.psych.55.090902.142050

Dum, R. P., & Strick, P. L. (2002). Motor areas in the frontal lobe of the primate. *Physiology & behavior, 77*(4), 677-682. doi:10.1016/S0031-9384(02)00929-0

Duncan, J. (1986). Disorganisation of behaviour after frontal lobe damage. *Cognitive Neuropsychology, 3*(3), 271-290. doi:10.1080/02643298608253360

Durbin, R., & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature, 343*(6259), 644-647.

Durstewitz, D., Vittoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron, 66*(3), 438-448. doi:10.1016/j.neuron.2010.03.029

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2007). Evolutionary development of hierarchical learning structures. *IEEE transactions on evolutionary computation, 11*(2), 249-264. doi:10.1109/tevc.2006.890270

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences, 17*(5), 202-209. doi:10.1016/j.tics.2013.03.006

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19-23. doi:10.1111/1467-8721.00160

Erwin, E., Obermayer, K., & Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural computation, 7*(3), 425-468. doi:10.1162/neco.1995.7.3.425

Evarts, E. V., Shinoda, Y., & Wise, S. P. (1984). *Neurophysiological approaches to higher brain functions*. Cambridge: John Wiley & Sons.

Everling, S., Tinsley, C. J., Gaffan, D., & Duncan, J. (2006). Selective representation of task-relevant objects and locations in the monkey prefrontal cortex. *European Journal of Neuroscience, 23*(8), 2197-2214. doi:10.1111/j.1460-9568.2006.04736.x

Fantino, E. (2003). Pigeon parallels to human metacognition. *Behavioral and Brain Sciences, 26*(3), 343-344. Retrieved from https://www.cambridge.org/core/journals/behavioral-and-brain-sciences

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory: updated edition*. Hillsdale, NJ: Erlbaum.

Feldman, J. A. (2016). Actionability and Simulation: No Representation without Communication. *Frontiers in Psychology, 7*, 1457. doi:10.3389/fpsyg.2016.01457

Feng, S. F. (2012). *Extensions and applications of stochastic accumulator models in attention and decision making.* (Doctor of Philosophy), Princeton University,

Feng, S. F., Schwemmer, M., Gershman, S., & Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience, 14*(1), 129-146. doi:10.3758/s13415-013-0236-9

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477-531.

Fodor, J. A. (1975). *The Language of Thought* (Vol. 5). Boston: Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge: MIT press.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*(2), 141-205. doi:10.1207/s15516709cog1902_1

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*(7084), 680-683. doi:10.1038/nature04587

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review, 113*(2), 300. doi:10.1037/0033-295X.113.2.300

Freeman, W. J. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences, 18*(4), 631-631. doi:10.1017/S0140525X0004022X

Frijda, N. H. (2016). The evolutionary emergence of what we call "emotions". *Cognition and Emotion, 30*(4), 609-620. doi:10.1080/02699931.2016.1145106

Fuster, J. M. (1988). Prefrontal cortex. In *Comparative neuroscience and neurobiology* (pp. 107-109). New York: Springer.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science, 173*(3997), 652-654.

Gabbay, D. M., Woods, J., & Thagard, P. (2006). *Philosophy of psychology and cognitive science*. Oxford: Elsevier.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493-501. doi:10.1016/S1364-6613(98)01262-5

Gallistel, C. R. (2006). The nature of learning and the functional architecture of the brain. *Psychological science around the world, 1*, 63-71.

Gallistel, C. R. (2008). Learning and representation. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference* (pp. 2-42). New York: Elsevier.

Gallistel, C. R. (2013). *The organization of action: A new synthesis*. New Jersey: Psychology Press.

Garcia, J., Ervin, F., & Koelling, R. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science, 5*(3), 121-122. doi:10.3758/BF03328311

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452-459.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*(5), 350-363. doi:10.1038/nrn3476

Giurfa, M. (2007). Behavioral and neural analysis of associative learning in the honeybee: a taste from the magic well. *Journal of Comparative Physiology A, 193*(8), 801-824. doi:10.1007/00359-007-0235-9

Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2002). Environmental complexity, signal detection, and the evolution of cognition. In M. Bekoff, C. Allen, & G. M. Burghardt (Eds.), *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition* (pp. 135-142). Cambridge: MIT Press.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience, 10*(5), 1563-1573. doi:10.1046/j.1460-9568.1998.00181.x

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review Neuroscience, 30*, 535-574. doi:10.1146/annurev.neuro.29.051605.113038

Goldman-Rakic, P. S. (1987). Development of cortical circuitry and cognitive function. *Child development, 58*(3), 601-622. doi:10.2307/1130201

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron, 14*(3), 477-485. doi:10.1016/0896-6273(95)90304-6

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences, 93*(24), 13473-13480. doi:http://www.pnas.org.simsrad.net.ocs.mq.edu.au/content/93/24/13473.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Comprehensive Physiology*. doi:10.1002/cphy.cp010509

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American psychologist, 54*(7), 493. doi:10.1037/0003-066X.54.7.493

Goodhill, G. J., & Willshaw, D. (1990). Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network: Computation in neural systems, 1*(1), 41-59. doi:10.1088/0954-898X_1_1_004

Gopnik, A., Schulz, L., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation*. London: Oxford University Press.

Graziano, M. S. (2008). *The intelligent movement machine: An ethological perspective on the primate motor system*. London: Oxford University Press.

Graziano, M. S. (2016). Ethological action maps: a paradigm shift for the motor cortex. *Trends in Cognitive Sciences, 20*(2), 121-132. doi:10.1016/j.tics.2015.10.008

Graziano, M. S., & Aflalo, T. N. (2007). Mapping behavioral repertoire onto the cortex. *Neuron, 56*(2), 239-251. doi:10.1016/j.neuron.2007.09.013

Graziano, M. S., Aflalo, T. N., & Cooke, D. F. (2005). Arm movements evoked by electrical stimulation in the motor cortex of monkeys. *Journal of Neurophysiology, 94*(6), 4209-4223. doi:10.1152/jn.01303.2004

Graziano, M. S., Cooke, D. F., Taylor, C. S., & Moore, T. (2004). Distribution of hand location in monkeys during spontaneous behavior. *Experimental Brain Research, 155*(1), 30-36. doi:10.1007/s00221-003-1701-4

Graziano, M. S., Taylor, C. S., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron, 34*(5), 841-851. doi:10.1016/S0896-6273(02)00698-0

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, 3*(2), 291-310. doi:10.1016/0010-0285(72)90009-6

Greenfield, P. M., & Schneider, L. (1977). Building a tree structure: The development of hierarchical complexity and interrupted strategies in children's construction activity. *Developmental Psychology, 13*(4), 299. doi:10.1037/0012-1649.13.4.299

Greenspan, R. (2007). *An introduction to nervous systems.* . Cold Spring Harbor: Cold Spring Harbor Press.

Gregory, R. L. (1969). *Eye and brain: the psychology of seeing*. New Jersey: Princeton University Press.

Gregory, R. L. (2005). The Medawar Lecture 2001 Knowledge for vision: vision for knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1458), 1231-1251. doi:10.1098/rstb.2005.1662

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357-364. doi:10.1016/j.tics.2010.05.004

Güntürkün, O. (2005). Avian and mammalian "prefrontal cortices": Limited degrees of freedom in the evolution of the neural mechanisms of goal-state maintenance. *Brain Research Bulletin, 66*(4–6), 311-316. doi:10.1016/j.brainresbull.2005.02.004

Gurney, K. N., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological cybernetics, 84*(6), 401-410. doi:10.1007/l00007984

Harrison, T. C., Ayling, O. G., & Murphy, T. H. (2012). Distinct cortical circuit mechanisms for complex forelimb movement and motor map topography. *Neuron, 74*(2), 397-409. doi:10.1016/j.neuron.2012.02.028

Hartner, D. F. (2013). Conceptual analysis as armchair psychology: in defense of methodological naturalism. *Philosophical studies, 165*(3), 921-937. doi:10.1007/s11098-012-9981-9

Haugeland, J. (1991). Representational genera. In W. Ramsey & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61-89).

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3*(4), 275-310. doi:10.1207/s15516709cog0304_1

Hazelbauer, G. L., Falke, J. J., & Parkinson, J. S. (2008). Bacterial chemoreceptors: high-performance signaling in networked arrays. *Trends in biochemical sciences, 33*(1), 9-19. doi:10.1016/j.tibs.2007.09.014

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1485), 1601-1613. doi:10.1098/rstb.2007.2055

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: John Wiley and Sons, Inc.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences, 114*(8), 1773-1782. doi:10.1073/pnas.1619788114

Herrnstein, R. J. (1990). Levels of stimulus control: A functional approach. *Cognition, 37*(1–2), 133-166. doi:10.1016/0010-0277(90)90021-B

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428-434. doi:10.1016/j.tics.2007.09.004

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed Representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition.* (Vol. 1). Cambridge: The MIT Press.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507. doi:10.1126/science.1127647

Hohwy, J. (2013). *The Predictive Mind*. London: Oxford University Press.

Holland, L. Z., & Holland, N. D. (1999). Chordate origins of the vertebrate central nervous system. *Current Opinion in Neurobiology, 9*(5), 596-602. doi:10.1016/S0959-4388(99)00003-3

Horton, J. C., & Adams, D. L. (2005). The cortical column: a structure without a function. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*(1456), 837-862. doi:10.1098/rstb.2005.1623

Hoshi, E., Shima, K., & Tanji, J. (1998). Task-Dependent Selectivity of Movement-Related Neuronal Activity in the Primate Prefrontal Cortex. *Journal of Neurophysiology, 80*(6), 3392-3397. Retrieved from http://jn.physiology.org/content/jn/80/6/3392.full.pdf

Hoy, R. R. (1989). Startle, categorical response, and attention in acoustic behavior of insects. *Annual Review of Neuroscience, 12*(1), 355-375. doi:10.1146/annurev.ne.12.030189.002035

Hoyer, P. O., & Hyvärinen, A. (2003). *Interpreting neural response variability as Monte Carlo sampling of the posterior.* Paper presented at the Advances in neural information processing systems.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology, 148*(3), 574-591. doi:10.1113/jphysiol.1959.sp006308

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Apple Century Crofts.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism : basic minds without content*. Cambridge: MIT Press.

Isogai, F., Kato, T., Fujimoto, M., Toi, S., Oka, A., Adachi, T., . . . Masuda, Y. (2012). Cortical area inducing chewing-like rhythmical jaw movements and its connections with thalamic

nuclei in guinea pigs. *Neuroscience research, 74*(3), 239-247. doi:10.1016/j.neures.2012.10.009

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. London: Oxford University Press.

Jackson, J. H. (1873). On the anatomical & physiological localisation of movements in the brain. *The Lancet, 101*(2581), 232-235. doi:10.1016/S0140-6736(02)63385-9

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), 187-202. doi:10.1017/S0140525X00034026

Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. London: Oxford University Press.

Jirenhed, D.-A., Rasmussen, A., Johansson, F., & Hesslow, G. (2017). Learned response sequences in cerebellar Purkinje cells. *Proceedings of the National Academy of Sciences, 114*(23), 6127-6132. doi:10.1073/pnas.1621132114

Kaas, J. H., & Catania, K. C. (2002). How do features of sensory representations develop? *Bioessays, 24*(4), 334-343. doi:10.1002/bies.10076

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (Vol. 1, pp. 29-61). New York: Academic Press.

Kalenscher, T., Güntürkün, O., Calabrese, P., Gehlen, W., Kalt, T., & Diekamp, B. (2005). Neural correlates of a default response in a delayed go/no-go task. *Journal of the Experimental Analysis of Behavior, 84*(3), 521-535. doi:10.1901/jeab.2005.86-04

Kalhat, J. (2015). Varieties of Representation. *Philosophy, 91*(1), 15-37. doi:10.1017/S0031819115000273

Kandel, E. R. (2009). The Biology of Memory: A Forty-Year Perspective. *The Journal of Neuroscience, 29*(41), 12748. doi:10.1523/jneurosci.3958-09.2009

Karlsson, M. P., Tervo, D. G., & Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science, 338*(6103), 135-139. doi:10.1126/science.1226518

Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge: MIT Press.

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science, 282*(5386), 108-111. doi:10.1126/science.282.5386.108

Katz, J. S., Wright, A. A., & Bodily, K. (2007). Issues in the comparative cognition of abstract-concept learning. *Comparative Cognition & Behavior Reviews, 2*, 79-92. doi:10.3819/ccbr.2008.20005

Katz, P. S., & Harris-Warrick, R. M. (1999). The evolution of neuronal circuits underlying species-specific behavior. *Current Opinion in Neurobiology, 9*(5), 628-633. doi:10.1016/S0959-4388(99)00012-4

Kawashima, R., O'Sullivan, B. T., & Roland, P. E. (1995). Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks: closing the" mind's eye". *Proceedings of the National Academy of Sciences, 92*(13), 5969-5972.

Keeler, J. D., Pichler, E. E., & Ross, J. (1989). Noise in neural networks: thresholds, hysteresis, and neuromodulation of signal-to-noise. *Proceedings of the National Academy of Sciences, 86*(5), 1712-1716.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. doi:10.1146/annurev.psych.55.090902.142005

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology, 10*(11), e1003915. doi:10.1371/journal.pcbi.1003915

Kis, A., Hernádi, A., Kanizsár, O., Gácsi, M., & Topál, J. (2015). Oxytocin induces positive expectations about ambivalent stimuli (cognitive bias) in dogs. *Hormones and behavior, 69*, 1-7. doi:10.1016/j.yhbeh.2014.12.004

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science, 302*(5648), 1181-1185. doi:10.1126/science.1088545

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics, 43*(1), 59-69. doi:10.1007/f00337288

Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *NeuroImage, 6*(4), 320-334. doi:10.1006/nimg.1997.0295

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences, 108*(33), 13852-13857. doi:10.1073/pnas.1101328108

Kriegeskorte, N. (2014). *The geometry of high-level visual representations.* Paper presented at the I-Perception, London.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine learning, 1*(1), 11-46. doi:10.1007/f00116249

Lake, B. M., Ullman, T., Tenenbaum, J., & Gershman, S. (2016). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences, 24*, 1-101. doi:10.1017/S0140525X16001837

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). *Deep Neural Networks Predict Category Typicality Ratings for Images.* Paper presented at the CogSci.

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica, 43*(3), 199-224. doi:10.1016/0001-6918(79)90026-X

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), (pp. 112-145). New York: Wiley.

Lattal, K. A. (2010). Delayed Reinforcement of Operant Behavior. *Journal of the Experimental Analysis of Behavior, 93*(1), 129-139. doi:10.1901/jeab.2010.93-129

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: core readings* (pp. 3-81). Cambridge: MIT Press.

Lazareva, O. F., & Wasserman, E. A. (2008). Categories and Concepts in Animals. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference* (Vol. 1, pp. 197-226). Oxford: Academic Press.

Lebrecht, S., Bar, M., Barrett, L., & Tarr, M. (2012). Micro-Valences: Perceiving Affective Valence in Everyday Objects. *Frontiers in Psychology, 3*(107). doi:10.3389/fpsyg.2012.00107

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi:10.1038/nature14539

Lennie, P. (2003). The cost of cortical computation. *Current Biology, 13*(6), 493-497. doi:10.1016/S0960-9822(03)00135-0

Lepora, N. F., & Gurney, K. N. (2012). The basal ganglia optimize decision making over general perceptual hypotheses. *Neural computation, 24*(11), 2924-2945. doi:10.1162/neco_a_00360

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027-1038. doi:10.1016/j.conb.2012.06.001

Lezak, M. D. (2004). *Neuropsychological assessment*. Cambridge: Oxford University Press.

London, M., & Häusser, M. (2005). Dendritic computation. *Annual Review of Neuroscience, 28*, 503-532. doi:10.1146/annurev.neuro.28.061604.135703

Luna, B., Padmanabhan, A., & O'Hearn, K. (2010). What has fMRI told us about the development of cognitive control through adolescence? *Brain and cognition, 72*(1), 101-113. doi:10.1016/j.bandc.2009.08.005

Luria, A. R. (1970). The functional organization of the brain. *Scientific American, 222*, 66-78.

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Frontiers in Microbiology, 6*(264). doi:10.3389/fmicb.2015.00264

Mackintosh, N. J. (2000). Abstraction and Discrimination. In C. M. Heyes & L. Huber (Eds.), *The Evolution of Cognition* (pp. 123-141). Cambridge: MIT Press.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 126-135. doi:10.1037/0278-7393.14.1.126

Macnab, R. M., & Koshland, D. (1972). The gradient-sensing mechanism in bacterial chemotaxis. *Proceedings of the National Academy of Sciences, 69*(9), 2509-2512.

Maes, E., De Filippo, G., Inkster, A., Lea, S., De Houwer, J., D'Hooge, R., . . . Wills, A. (2015). Feature- versus rule-based generalization in rats, pigeons and humans. *ANIMAL COGNITION, 18*(6), 1267-1284. doi:10.1007/s10071-015-0895-8

Maes, P. (1993). Modeling adaptive autonomous agents. *Artificial life, 1*(1_2), 135-162. doi:10.1162/artl.1993.1.1_2.135

Manger, P. R., Woods, T. M., Muñoz, A., & Jones, E. G. (1997). Hand/face border as a limiting boundary in the body representation in monkey somatosensory cortex. *Journal of Neuroscience, 17*(16), 6338-6351.

Marder, E., & Thirumalai, V. (2002). Cellular, synaptic and network effects of neuromodulation. *Neural Networks, 15*(4), 479-493. doi:10.1016/S0893-6080(02)00043-6

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge: MIT Press.

Marshall, J. A., Bogacz, R., & Gilchrist, I. D. (2012). Consistent implementation of decisions in the brain. *PloS one, 7*(9), e43443. doi:10.1371/journal.pone.0043443

Marshall, L., Mathys, C., Ruge, D., de Berker, A. O., Dayan, P., Stephan, K. E., & Bestmann, S. (2016). Pharmacological fingerprints of contextual uncertainty. *PLoS biology, 14*(11), e1002575. doi:10.1371/journal.pbio.1002575

Martinetz, T. (1993). Competitive Hebbian learning rule forms perfectly topology preserving maps. In *ICANN'93* (pp. 427-434): Springer.

Mather, G. (2016). *Foundations of sensation and perception*: Psychology Press.

McClelland, J. L. (1988). *Parallel distributed processing: Implications for cognition and development*. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA219063

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14*(8), 348-356. doi:10.1016/j.tics.2010.06.002

McGinley, M. J., David, S. V., & McCormick, D. A. (2015). Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron, 87*(1), 179-192. doi:10.1016/j.neuron.2015.05.038

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., & Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends in Neurosciences, 28*(8), 401-407. doi:10.1016/j.tins.2005.06.006

Menzel, R. (2013). Learning, Memory, and Cognition: Animal Perspectives. In C. G. Galizia & P.-M. Lledo (Eds.), *Neurosciences* (pp. 629-653). Berlin: Springer-Verlag.

Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., . . . Schaupp, F. (2005). Honey bees navigate according to a map-like spatial memory. *Proceedings of the National Academy of Sciences of the United States of America, 102*(8), 3040-3045. doi:10.1073/pnas.0408550102

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and Hierarchically Modular Organization of Brain Networks. *Frontiers in Neuroscience, 4*, 200. doi:10.3389/fnins.2010.00200

Miller, E. K. (2000). The prefontral cortex and cognitive control. *Nature Reviews Neuroscience, 1*(1), 59-65. doi:10.1038/35036228

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167-202. doi:10.1146/annurev.neuro.24.1.167

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *The Journal of Neuroscience, 16*(16), 5154-5167. Retrieved from www.jneurosci.org/content/16/16/5154.full.pdf+htm

Miller, E. K., Nieder, A., Freedman, D. J., & Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology, 13*(2), 198-203. doi:10.1016/S0959-4388(03)00037-0

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences, 7*(3), 141-144. doi:10.1016/S1364-6613(03)00029-9

Miller, G. A., Galanter, E., & Pribram, K. H. (1986). *Plans and the structure of behavior*. New York: Adams Bannister Cox.

Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives, 9*, 185-200. doi:10.2307/2214217

Minors, D. (2016). *How a honey bee solves a conceptual problem.* (Bachelor of Psychology (Honours)), Macquarie University, Sydney.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology, 41*(1), 49-100. doi:10.1006/cogp.1999.0734

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529-533. doi:10.1038/nature14236

Mockus, J. (2012). *Bayesian approach to global optimization: theory and applications* (Vol. 37). Boston: Kluwer Academic Publishers.

Moore, T. L., Schettler, S. P., Killiany, R. J., Rosene, D. L., & Moss, M. B. (2012). Impairment in Delayed Non-Matching to Sample Following Lesions of Dorsal Prefrontal Cortex. *Behavioral Neuroscience, 126*(6), 772-780. doi:10.1037/a0030493

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences, 108*(30), 12491-12496.

Moser, M.-B., Rowland, D. C., & Moser, E. I. (2015). Place cells, grid cells, and memory. *Cold Spring Harbor perspectives in biology, 7*(2), a021808.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain: a journal of neurology, 120*(4), 701-722. doi:10.1093/brain/120.4.701

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science, 17*(4), 463-496. doi:10.1207/s15516709cog1704_1

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113-126. doi:10.1145/360018.360022

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133. doi:10.1103/PhysRevE.69.066133

Niell, C. M., & Stryker, M. P. (2008). Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience, 28*(30), 7520-7536.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience, 12*(2), 241-268. doi:10.3758/s13415-011-0083-5

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139-154. doi:10.1016/j.jmp.2008.12.005

Norman, D. A., & Shallice, T. (1980). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation* (pp. 1-18). New York: Plenum Press.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation, 18*(2), 283-328.

Obermayer, K., Blasdel, G. G., & Schulten, K. (1992). Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Physical Review A, 45*(10), 7568. doi:10.1103/PhysRevA.45.7568

Obermayer, K., Ritter, H., & Schulten, K. J. (1992). A model for the development of the spatial structure of retinotopic maps and orientation columns. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 75*(5), 537-545.

Obermayer, K., Schulten, K., & Blasdel, G. (1991). *A neural network model for the formation and for the spatial structure of retinotopic maps, orientation-and ocular dominance columns*: University of Illinois at Urbana-Champaign.

Owen, A. M., Sahakian, B. J., Semple, J., Polkey, C. E., & Robbins, T. W. (1995). Visuo-spatial short-term recognition memory and learning after temporal lobe excisions, frontal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia, 33*(1), 1-24. doi:10.1016/0028-3932(94)00098-a

Pacherie, É. (2002). Emotion and action. *European Review of Philosophy, 5*, 55-90.

Pais, D., Hogan, P. M., Schlegel, T., Franks, N. R., Leonard, N. E., & Marshall, J. A. (2013). A mechanism for value-sensitive decision-making. *PloS one, 8*(9), e73216. doi:10.1371/journal.pone.0073216

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. London: Oxford University Press.

Parker, A. J., & Newsome, W. T. (1998). Sense and the Single Neuron: Probing the Physiology of Perception. *Annual Review of Neuroscience, 21*(1), 227. doi:10.1146/annurev.neuro.21.1.227

Passingham, R. E. (1993). *The frontal lobes and voluntary action*. New York: Oxford University Press.

Peil, K. T. (2014). Emotion: the self-regulatory sense. *Global Advances in Health and Medicine, 3*(2), 80-108. doi:10.7453/gahmj.2013.058

Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man; a clinical study of localization of function*. New York: Macmillan.

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry, 37*(1), 51-87. doi:10.1111/j.1469-7610.1996.tb01380.x

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an African Grey parrot (Psittacus erithacus): Learning with respect to categories of color, shape, and material. *Animal Learning & Behavior, 15*(4), 423-432. doi:10.3758/bf03205051

Perry, C. J., & Barron, A. B. (2013). Neural Mechanisms of Reward in Insects. *Annual Review of Entomology, 58*, 543-562. doi:10.1146/annurev-ento-120811-153631

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). *Adapting deep network features to capture psychological representations.* Paper presented at the Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia.

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal-and temporal-lobe lesions in man. *Neuropsychologia, 20*(3), 249-262. doi:10.1016/0028-3932(82)90100-2

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place cell sequences depict future paths to remembered goals. *Nature, 497*(7447), 74-79. doi:10.1038/nature12112

Pirrone, A., Stafford, T., & Marshall, J. A. (2014). When natural selection should optimize speed-accuracy trade-offs. *Frontiers in Neuroscience, 8*, 73. doi:10.3389/fnins.2014.00073

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews, 32*(2), 197-205. doi:10.1016/j.neubiorev.2007.07.007

Pomerantz, J. R. (2003). Perception: Overview. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (Vol. 3, pp. 527–537). London: Nature Pub. Group.

Porter, R. (1985). Neurophysiological approaches to higher brain functions. *Science, 228*, 1421-1421. doi:10.1126/science.228.4706.1421

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In D. A. Balota & E. J. Marsh (Eds.), *Cognitive psychology: Key readings* (pp. 205-223). New York: Psychology Press.

Premack, D. (1978). On the abstractness of human concepts: Why it would be difficult to talk to a pigeon. In S. H. Hulse, H. Fowler, & W. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 423-451). Hillsdale: Erlbaum.

Premack, D., & Premack, A. J. (1983). *The mind of an ape*. New York: W.W. Norton.

Prescott, T. J., Redgrave, P., & Gurney, K. N. (1999). Layered control architectures in robots and vertebrates. *Adaptive Behavior, 7*(1), 99-127. doi:10.1177/105971239900700105

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2658-2663. doi:10.1073/pnas.0400054101

Ramanathan, D., Conner, J. M., & Tuszynski, M. H. (2006). A form of motor cortical plasticity that correlates with recovery of function after brain injury. *Proceedings of the National Academy of Sciences, 103*(30), 11370-11375. doi:10.1073/pnas.0601065103

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology, 20*(2), 262-270. doi:10.1016/j.conb.2010.03.001

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59.

Ravizza, S. M., & Carter, C. S. (2008). Shifting set about task switching: behavioral and neural evidence for distinct forms of cognitive flexibility. *Neuropsychologia, 46*(12), 2924-2935. doi:10.1016/j.neuropsychologia.2008.06.006

Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience, 15*(4), 574-583. doi:10.1162/089892903321662958

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience, 89*(4), 1009-1023. doi:10.1016/S0306-4522(98)00319-4

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review, 74*(3), 151. doi:10.1037/h0024475

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge: MIT Press.

Roitblat, H. L., & von Fersen, L. (1992). Comparative cognition: Representations and processes in learning and memory. *Annual Review of Psychology, 43*(1), 671-710.

Rosa, M. G. (2002). Visual maps in the adult primate cerebral cortex: some implications for brain development and evolution. *Brazilian Journal of Medical and Biological Research, 35*(12), 1485-1498. doi:10.1590/S0100-879X2002001200008

Rosa, M. G., & Schmid, L. M. (1995). Visual areas in the dorsal and medial extrastriate cortices of the marmoset. *Journal of Comparative Neurology, 359*(2), 272-299. doi:10.1002/cne.903590207

Rosa, M. G., & Tweedale, R. (2005). Brain maps, great and small: lessons from comparative studies of primate visual cortical organization. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 665-691. doi:10.1098/rstb.2005.1626

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America, 102*(20), 7338-7343. doi:10.1073/pnas.0502455102

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructures of cognition.* (Vol. 1). Cambridge: The MIT Press.

Rustichini, A. (2008). Neuroeconomics: formal models of decision making and cognitive neuroscience. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 33-46). New York: Elsevier.

Saarinen, J., & Kohonen, T. (1985). Self-organized formation of colour maps in a model cortex. *Perception, 14*(6), 711-719. doi:10.1068/p140711

Sacerdoti, E. D. (1974). Planning in a hierarchy of abstraction spaces. *Artificial intelligence, 5*(2), 115-135. doi:10.1016/0004-3702(74)90026-5

Said, C. P., & Heeger, D. J. (2013). A model of binocular rivalry and cross-orientation suppression. *PLOS Computational Biology, 9*(3), e1002991. doi:10.1371/journal.pcbi.1002991

Sarter, M., & Bruno, J. P. (1997). Cognitive functions of cortical acetylcholine: toward a unifying hypothesis. *Brain Research Reviews, 23*(1), 28-46. doi:10.1016/S0165-0173(96)00009-4

Schacter, D. L. (1996). *Searching for memory: The brain, the mind, and the past*. New York: Basic Books.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience, 2*(1), 33-42. doi:10.1038/35049054

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience, 16*(4), 486-492. doi:10.1038/nn.3331

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26*(1), 3-8. doi:10.1002/hipo.22523

Schieber, M. H. (2001). Constraints on somatotopic organization in the primary motor cortex. *Journal of Neurophysiology, 86*(5), 2125-2143.

Schiffer, A.-M., Siletti, K., Waszak, F., & Yeung, N. (2017). Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *NeuroImage, 146*, 626-641. doi:10.1016/j.neuroimage.2016.08.057

Schoenbaum, G., & Setlow, B. (2001). Integrating orbitofrontal cortex into prefrontal theory: common processing themes across species and subdivisions. *Learning & Memory, 8,* 134.

Schott, G. D. (1993). Penfield's homunculus: a note on cerebral cartography. *Journal of Neurology, Neurosurgery & Psychiatry, 56*(4), 329-333. doi:10.1136/jnnp.56.4.329

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience, 17*(3), 183-195. doi:10.1038/nrn.2015.26

http://www.nature.com/nrn/journal/v17/n3/abs/nrn.2015.26.html#supplementary-information

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593-1599. doi:10.1126/science.275.5306.1593

Schwaerzel, M., Monastirioti, M., Scholz, H., Friggi-Grelin, F., Birman, S., & Heisenberg, M. (2003). Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in Drosophila. *Journal of Neuroscience, 23*(33), 10495-10502.

Scott, S. H. (2006). Neuroscience: Converting thoughts into action. *Nature, 442*(7099), 141-142. doi:10.1038/442141a

Seeley, T. D., Visscher, P. K., Schlegel, T., Hogan, P. M., Franks, N. R., & Marshall, J. A. (2012). Stop signals provide cross inhibition in collective decision-making by honeybee swarms. *Science, 335*(6064), 108-111. doi:10.1126/science.1210361

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annual Review of Neuroscience, 33*(1), 203-219. doi:10.1146/annurev.neuro.051508.135546

Seligman, M. E. (1971). Phobias and preparedness. *Behavior therapy, 2*(3), 307-320. doi:10.1016/S0005-7894(71)80064-3

Seligman, M. E., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science, 8*(2), 119-141. doi:10.1177/1745691612474317

Semon, R. W. (1921). *The mneme*. London: G. Allen & Unwin Limited.

Seward, J. P. (1949). An experimental analysis of latent learning. *Journal of experimental psychology, 39*(2), 177-186. doi:10.1037/h0063169

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology, 86*(4), 1916-1936.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 298*(1089), 199-209.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*(2), 127-190. doi:10.1037/0033-295X.84.2.127

Shu, Y., Hasenstaub, A., & McCormick, D. A. (2003). Turning on and off recurrent balanced cortical activity. *Nature, 423*(6937), 288-293.

Skinner, B. F. (1938). *The behaviour of organisms: An experimental analysis*. Oxford: Appleton-Century.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience & Biobehavioral Reviews, 32*(2), 249-264. doi:10.1016/j.neubiorev.2007.07.009

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*(5408), 1657-1661. doi:10.1126/science.283.5408.1657

Soltoggio, A., Durr, P., Mattiussi, C., & Floreano, D. (2007). *Evolving neuromodulatory topologies for reinforcement learning-like problems.* Paper presented at the IEEE Congress on Evolutionary Computation, Singapore.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review, 119*(1), 120-154. doi:10.1037/a0026435

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences, 112*(37), 11708-11713. doi:10.1073/pnas.1505483112

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLOS Computational Biology, 10*(8), e1003779. doi:10.1371/journal.pcbi.1003779

Søvik, E., Perry, C. J., & Barron, A. B. (2015). Insect Reward Systems: Comparing Flies and Bees. *Advances in Insect Physiology, 48*, 189-226. doi:10.1016/bs.aiip.2014.12.006

Spencer, R. M., & Ivry, R. B. (2013). Cerebellum and timing. In M. Manto, J. D. Schmahmann, F. Rossi, D. L. Gruol, & N. Koibuchi (Eds.), *Handbook of the cerebellum and cerebellar disorders* (pp. 1201-1219). Netherlands: Springer.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*(3), 251-260. doi:10.1007/BF02289729

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643-662. doi:10.1037/h0054651

Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., . . . Izukawa, D. (2000). Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia, 38*(4), 388-402. doi:10.1016/S0028-3932(99)00093-7

Stuss, D. T., Shallice, T., Alexander, M. P., & Picton, T. W. (1995). A multidisciplinary approach to anterior attentional functions. *Annals of the New York Academy of Sciences, 769*(1), 191-212. doi:10.1111/j.1749-6632.1995.tb38140.x

Sutton, R. S., & Barto, A. G. (1990). Time Derivate Models of Pavlovian Reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (Vol. 12, pp. 497-537). Cambridge: MIT Press.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors, 45*(1), 61-76. doi:10.1518/hfes.45.1.61.27224

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research, 10*(Jul), 1633-1685.

Thomas, R. K. (2012). Learning set formation and conceptualization. In *Encyclopedia of the Sciences of Learning* (pp. 1966-1968): Springer.

Thompson, R. K. R., & Oden, D. L. (1995). A profound disparity revisited: Perception and judgment of abstract identity relations by chimpanzees, human infants, and monkeys. *Behavioural Processes, 35*(1), 149-161. doi:10.1016/0376-6357(95)00048-8

Thompson, R. K. R., & Oden, D. L. (2000). Categorical Perception and Conceptual Judgments by Nonhuman Primates: The Paleological Monkey and the Analogical Ape. *Cognitive Science, 24*(3), 363-396. doi:10.1207/s15516709cog2403_2

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. London: Macmillan.

Tolman, E. C. (1951). *Purposive behavior in animals and men*. London: Cambridge University Press.

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California publications in psychology, 4*, 257-275.

Trimmer, P. C., Houston, A. I., Marshall, J. A., Bogacz, R., Paul, E. S., Mendl, M. T., & McNamara, J. M. (2008). Mammalian choices: combining fast-but-inaccurate and slow-but-accurate decision-making systems. *Proceedings of the Royal Society of London B: Biological Sciences, 275*(1649), 2353-2361. doi:10.1098/rspb.2008.0417

Tsotsos, J. (2017). Attention and Cognition: Principles to Guide Modeling. In Q. Zhao (Ed.), *Computational and Cognitive Neuroscience of Vision* (pp. 277-295). Singapore: Springer.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence, 78*(1-2), 507-545. doi:10.1016/0004-3702(95)00025-9

Tsujimoto, S., Genovesio, A., & Wise, S. P. (2011). Comparison of strategy signals in the dorsolateral and orbital prefrontal cortex. *The Journal of Neuroscience, 31*(12), 4583. doi:10.1523/jneurosci.5816-10.2011

Usher, M., Elhalal, A., & McClelland, J. L. (2008). The neurodynamics of choice, value-based decisions, and preference reversal. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 277-300): Oxford Scholarship Online.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review, 108*(3), 550-592. doi:10.1037/0033-295X.108.3.550

Vergoz, V., Roussel, E., Sandoz, J.-C., & Giurfa, M. (2007). Aversive learning in honeybees revealed by the olfactory conditioning of the sting extension reflex. *PloS one, 2*(3), e288. doi:10.1371/journal.pone.0000288

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*(1), 37-58. doi:10.1080/00140137008931117

Wadhams, G. H., & Armitage, J. P. (2004). Making sense of it all: bacterial chemotaxis. *Nat Rev Mol Cell Biol, 5*(12), 1024-1037. doi:10.1038/nrm1524

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature, 411*(6840), 953-956. doi:10.1038/35082081

Wallis, J. D., & Miller, E. K. (2003). From Rule to Response: Neuronal Processes in the Premotor and Prefrontal Cortex. *Journal of Neurophysiology, 90*(3), 1790-1806. doi:10.1152/jn.00086.2003

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron, 36*(5), 955-968. doi:10.1016/S0896-6273(02)01092-9

Webb, B. (2004). Neural mechanisms for prediction: do insects have forward models? *Trends in Neurosciences, 27*(5), 278-282. doi:10.1016/j.tins.2004.03.004

Webb, B. (2006). Transformation, encoding and representation. *Current Biology, 16*(6), R184-R185. doi:10.1016/j.cub.2006.02.034

Webb, B. (2012). Cognition in insects. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1603), 2715. doi:10.1098/rstb.2012.0218

Whippo, C. W., & Hangarter, R. P. (2006). Phototropism: Bending towards Enlightenment. *The Plant Cell, 18*(5), 1110-1119. doi:10.1105/tpc.105.039669

White, M. I., & Wise, P. S. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research, 126*(3), 315-335. doi:10.1007/s002210050740

Wilson, H. R., Krupa, B., & Wilkinson, F. (2000). Dynamics of perceptual oscillations in form vision. *Nature neuroscience, 3*(2), 170-176. doi:10.1038/72115

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science, 14*(3), 131-134. doi:10.1111/j.0963-7214.2005.00355.x

Windhorst, U. (1996). On the role of recurrent inhibitory feedback in motor control. *Progress in neurobiology, 49*(6), 517-587. doi:10.1016/0301-0082(96)00023-8

Wirsig, C. R., & Grill, H. J. (1982). Contribution of the rat's neocortex to ingestive control: I. Latent learning for the taste of sodium chloride. *Journal of comparative and physiological psychology, 96*(4), 615-627. doi:10.1037/h0077911

Woolsey, C. N. (1952). Patterns of localization in sensory and motor areas of the cerebral cortex. In M. M. Fund (Ed.), *The biology of mental health and disease* (pp. 193-206). New York: Hoeber.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619-8624. doi:10.1073/pnas.1403112111

Yoganarasimha, D., Yu, X., & Knierim, J. J. (2006). Head direction cell representations maintain internal coherence during conflicting proximal and distal cue rotations: comparison with hippocampal place cells. *Journal of Neuroscience, 26*(2), 622-631. doi:10.1523/jneurosci.3885-05.2006

Yu, A. J., & Dayan, P. (2005). *Inference, attention, and decision in a Bayesian neural architecture.* Paper presented at the Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308. doi:10.1016/j.tics.2006.05.002

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience, 23*(12), 4057-4066. doi:10.1162/jocn_a_00078

Zayan, R., & Vauclair, J. (1998). Categories as paradigms for comparative cognition. *Behavioural Processes, 42*(2–3), 87-99. doi:10.1016/S0376-6357(97)00064-8

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *Journal of Neuroscience, 28*(49), 13194-13201. doi:10.1523/jneurosci.2915-08.2008

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of general psychology, 1*(2), 198-226. doi:10.1037/1089-2680.1.2.198

Zentall, T. R., Galizio, M., & Critchfield, T. S. (2002). Categorization, concept learning, and behaviour analysis: an introduction. *Journal of the Experimental Analysis of Behavior, 78*(3), 237-248. doi:10.1901/jeab.2002.78-237

Zentall, T. R., Wasserman, E. A., & Urcuioli, P. J. (2014). Associative concept learning in animals. *Journal of the Experimental Analysis of Behavior, 101*(1), 130-151. doi:10.1002/jeab.55