Section I

The Basics of Cognitive Control

Theoretical Constructs and Behavioural Phenomena

1

Cognitive Control

Core Constructs and Current Considerations

Jonathan D. Cohen

The capacity for cognitive control is perhaps the most distinguishing characteristic of human behaviour. Broadly defined, cognitive control refers to the ability to pursue goal-directed behaviour, in the face of otherwise more habitual or immediately compelling behaviours. This ability is engaged by every faculty that distinguishes human abilities from those of other species, and in virtually every domain of human function from perception to action, decision making to planning, and problem solving to language processing. Understanding the mechanisms that underlie our capacity for cognitive control seems essential to unravelling the mystery of why, on the one hand, we are capable of intelligent, goal-directed behaviour, whereas on the other hand this ability seems so vulnerable to irrational influences and failure. Not surprisingly, the distinction between controlled and automatic processing is one of the most fundamental and long-standing principles of cognitive psychology. However, as fundamental as the construct of cognitive control is, it has been almost equally elusive. Most importantly, the construct on its own says little about the mechanisms involved.

Fortunately, in the half-century since the concept of control was first introduced into psychology (Miller, Galanter, & Pribram, 1960), and afforded a central role in cognitive psychology not long thereafter (Posner & Snyder, 1975; Shiffrin & Schneider, 1977), considerable progress has been made in characterising cognitive control in more precise and mechanistic terms, at both the psychological and neurobiological levels of analysis (e.g., Anderson, 1983; Botvinick & Cohen, 2014; Collins & Frank, 2013; Daw, Niv, & Dayan, 2005; Dayan, 2012; Duncan, 2010; Koechlin & Summerfield, 2007; Miller & Cohen, 2001; O'Reilly, 2006). Much of this progress is reflected in the chapters of this volume. Needless to say, however, considerable progress remains to be made, and formulating a way forward may benefit by revisiting, and carefully reconsidering some of the foundational ideas that originally motivated the construct of cognitive control, and how these have evolved over the past half-century.

In this introduction, I review the original formulations of the distinction between controlled and automatic processing, the issues that this distinction raised, how these have been addressed, and questions that remain. This review is intended to be useful in at least two ways: (a) as a guide to the central constructs and most pressing issues concerning cognitive control for those who are new to this area of research; and (b) as an inventory of challenges that a satisfying account of cognitive control must address for those who are familiar with the area. I have organised the issues into three broad categories: (a) core, defining features of cognitive control; (b) the relationship of cognitive control to other closely related constructs

The Wiley Handbook of Cognitive Control, First Edition. Edited by Tobias Egner.

in psychology; and (c) ways in which the understanding of cognitive control may be informed by theoretical approaches that have proved valuable in other areas such as computer science.

Core Constructs

Definitional Attributes of Controlled Versus Automatic Processing

As a theoretical construct, cognitive control grew out the study of communications and control systems, including the discipline of cybernetics that flourished in the middle of the last century (e.g., Wiener, 1948). In particular, an influential book by Miller et al. (1960) explicitly drew the connection between control theory, the goal-directed nature of human cognition, and its apparently hierarchical structure—topics that have regained attention in modern research (as will be discussed below). However, three articles are generally credited with operationalising the construct of cognitive control, and placing it at the centre of experimental research in cognitive psychology: one by Posner & Snyder (1975), and a pair by Shiffrin and Schneider (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). These articles focused on three attributes that distinguish controlled from automatic processes: (a) controlled processes are slower to execute; (b) are subject to interference by competing automatic processes; (c) and rely on a central, limited-capacity processing mechanism.

The canonical example chosen by Posner & Snyder (1975) to illustrate this was a comparison of colour naming and word reading in the Stroop task (MacLeod, 1991; Stroop, 1935). Adults are almost universally faster to read a word out loud than to name the colour of a stimulus (criterion 1). Critically, when responding to incongruent stimuli (e.g., the word 'RED' displayed in green), the colour of the stimulus has almost no impact on the word reading response, whereas the word invariably interferes with naming the colour. Furthermore, attempts to name the colour while performing another unrelated task (such as mental arithmetic) are likely to be impaired. These properties generally do not apply to word reading. These findings were explained by proposing that colour naming is a controlled process, whereas word reading is automatic. This account of findings in the Stroop task quickly became—and in many areas still remains—a foundational paradigm for studying controlled and automatic processing (for example, the same principles are used to infer the influence of automatic processes using the Implicit Association Task—IAT; Greenwald & Banaji, 1995). However, almost as soon as the construct of controlled processing was introduced, it raised concerns.

Capacity Constraints

Central, limited-capacity mechanism. Perhaps the most important and controversial assertion was that cognitive control relies on a central, limited-capacity processing mechanism that imposes a serial constraint on the execution of controlled processes, as distinct from automatic processes that can be carried out in parallel.[1] The importance of this assumption cannot be overestimated. The idea was paradigmatic in the literal sense. It provided the operational criterion that is used almost universally to identify a process as control demanding: dual-task interference. If performance of a task suffers when another task that is unrelated (i.e., does not involve the same stimuli or responses) must be performed at the same time, then the first task is deemed to be control demanding. However, as practically—and introspectively—appealing as this assertion is, it is equally problematic.

The capacity constraints on control are generally attributed to its reliance on a limited resource; however, neither the nature of the resource, nor the reason for its limitation has

yet been identified. Some have argued that the resource may be metabolic (see discussion of effort below). However, there is no reliable evidentiary basis for this (Carter, Kofler, Forster, & McCullough, 2015), and it seems improbable given the importance of the function, the metabolic resources available to the brain, and the scale on which it is able to commit metabolic resources to other processes.

Another suggestion is that the limitation is structural. For example, most models of cognitive control propose that control-demanding processes rely on the activation and maintenance of control representations that are used to guide execution (see discussion below). These representations are considered the 'resource' upon which control relies, and the limited capacity of control is attributed to a limitation in the scope of such representations that can be actively maintained (e.g., a limitation in working memory capacity). However, this begs the question: Why is *that* capacity limited? One possibility is a physical limitation (akin to the limited number of memory registers in a CPU). However, like metabolic constraints, this seems highly improbable. There are 100 billion neurons in the human brain, of which about one third reside in areas thought to be responsible for cognitive control (e.g., the prefrontal and dorsal parietal cortex). With those resources available, evolution would have to be viewed as a poor engineer to be incapable of maintaining more than a single control representation at a time. Another possibility is that there are functional constraints on the system; for example, the number of representations that can be simultaneously maintained in an attractor system, or a tension between their number and resolution (Edin et al., 2009; Ma & Huang, 2009; Usher, Cohen, Haarmann, & Horn, 2001). Such efforts reflect important progress being made in developing quantitative, mechanistically explicit accounts of representation and processing in neural systems, and may well explain constraints within circumscribed domains of processing. However, once again, this begs the question: Why cannot a system as vast as the human neocortex proliferate attractor systems for a function as valuable as cognitive control?

Multiple resources hypothesis. An alternative to the idea that dual-task interference reflects a constraint in the control system *itself* is the idea that, instead, it reflects something about the processes being controlled. This idea has its origins in multiple resource theories of attention (Allport, 1980; Logan, 1985; Navon & Gopher, 1979; also see Allport, Antonis, & Reynolds, 1972; Wickens, 1984). They argued that interference between tasks may reflect cross-talk within *local* resources (e.g., representations or processes) needed to perform different tasks if they must make simultaneous use of those resources for different purposes—a problem that can arise anywhere in the system, and not just within the control system itself. A classic example of such cross-talk (Shaffer, 1975) contrasted two dual-task conditions: repeating an auditory stream ('echoing') while simultaneously typing visually presented text (copy-typing), versus simultaneously reading aloud and taking dictation. The former pair is relatively easy to learn, while the latter is considerably more difficult. The multiple resources explanation suggests that echoing and copy-typing involve non-overlapping local representations and processing pathways (one auditory—phonological—verbal, and the other visual—orthographic—manual). Because they make use of distinct resources there is no risk of cross-talk, and so it is possible to do both at once. In contrast, reading out loud and taking dictation make dual competing use of phonological representations (e.g., the one to be read and the one to be transcribed), and similarly for orthographic representations, and thus are subject to the problem of cross-talk.

This idea has been expressed in a number of models addressing cognitive control (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Cohen, Dunbar, & McClelland, 1990; Meyer & Kieras, 1997; Salvucci & Taatgen, 2008). These models suggest that constraints on the simultaneous execution of multiple tasks can be viewed as the *purpose* of control, rather than a limitation in its ability: A process relies on control whenever it risks coming into conflict

with (i.e., is subject to cross-talk from) another process. Such conflict can impair performance (either by slowing processing, or generating overt errors). Thus, a critical function of the control system is to monitor for such indicators (e.g., Botvinick et al., 2001; Holroyd & Coles, 2002), in order to limit processing in such a way as to avoid such conflicts—in just the way that traffic signals are meant to limit collisions among vehicles travelling on intersecting thoroughfares. There is considerable empirical evidence in support of this (e.g., Carter et al., 1998; Egner & Hirsch, 2005; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; Venkatraman, Rosati, Taren, & Huettel, 2009; Yeung, Botvinick, & Cohen, 2004). Recent computational work has suggested that even modest amounts of overlap among processing streams within a system can impose surprisingly strict limitations on the number of processes that can be safely executed at a given time (Feng, Schwemmer, Gershman, & Cohen, 2014; Musslick et al., 2016). Furthermore, these restrictions can be nearly scale invariant, and thus may offer a plausible account for the strikingly strict constraints on control-dependent behaviour in a system as computationally rich as the human brain.

Constraints on the ability to sustain control. Although most discussions of limited capacity focus on a numerical constraint (that is, how *many* control-dependent processes can be executed at once), there is an equally impressive and consistently observed *temporal* constraint (how *long* control can be sustained for a given task). Here again, the constraint has been assumed to reflect a limited resource. One popular version of this account—the 'ego depletion' hypothesis (Baumeister, Bratslavsky, Muraven, & Tice, 1998)—proposes that the resource is literally energetic, and the inability to sustain control reflects metabolic fatigue. Although this concurs with the subjective sense of effort associated with control, recent studies have called into question both the physiological basis (Kurzban, 2010) and empirical support (Carter et al., 2015) for this hypothesis. An alternative is that temporal constraints may reflect motivational factors (Inzlicht & Schmeichel, 2012). For example, as discussed in the section that follows, effort may reflect the signalling of opportunity costs associated with persistent performance of a given task, rather than a metabolic expense.

Effort and Motivation

From the earliest formulations, the construct of cognitive control has been closely associated with effort and motivation, an association that continues to be a focus of modern research (as evidenced by Chapters 9, 10, 23, and 24 in this volume on the topic by Chiew & Braver, Kool et al., Winecoff & Huettel, and Krebs & Woldorff). These terms have frequently elicited concern. An obvious and persistent one is about the qualitative—and potentially irreducibly subjective—nature of the phenomena to which they refer. Another is the awkwardness of the fit, perhaps most notably with regard to the original example: For example, is it really any more *effortful* simply to name the colour of an object, than it is to read a word? Despite these concerns, there are at least two reasons for considering effort and motivation, and their association with control.

Phenomenologically, the experience of effort helps identify and characterise conditions that seem to engage control. For example, while it may not be particularly effortful to name the colour of an apple, it *is* effortful to name the colour of an incongruent stimulus in the Stroop task. What is the difference? In one case, there is no interference, while in the other there is. Thus, the experience of effort—though subjective—offers a clue that control is engaged by conflict between processes (or the potential thereof), rather than by a particular process itself. This aligns with the idea, noted above, that a fundamental function of control is to reduce interference where it can arise.

The association of effort and motivation with control also highlights a theoretical construct that, until recently, has been all but ignored in research on control: the *cost* of control.

Irrespective of the sources of constraint, control mechanisms must manage a limited budget, with the consequence that allocating control to one process incurs an opportunity cost for others. This is something that was recognised in the earliest conceptions of cognitive control (Shiffrin & Schneider, 1977). From this perspective, 'effort' might be viewed as the phenomenological correlate of a signal indexing the opportunity cost associated with the allocation of control, and 'motivation' as the system's 'willingness to pay' the cost of control. These ideas are gaining currency in modern research, as reflected by several chapters in this volume (see also Braver, 2015; Kurzban, Duckworth, Kable, & Myers, 2013), and represent an exciting and important direction for future research. In particular, it offers the promise of bringing research on the mechanisms underlying cognitive control into contact with the study of value-based decision making and disturbances of behaviour specifically related to valuation and motivation, such as savings, drug addiction, and gambling (e.g., Bickel, Jarmolowicz, Mueller, Gatchalian, & McClure, 2012; Westbrook & Braver, 2015; Chapter 23 by Winecoff & Huettel in this volume).

The Continuum of Automaticity and Control

Context dependence. Another major concern with early formulations of controlled versus automatic processing was about the treatment of this distinction as a dichotomy. This quickly met with empirical challenges (Kahneman & Treisman, 1984). In one particularly striking example, each in a set of arbitrary shapes was assigned a colour word as its name, and participants were taught to name the shapes (MacLeod & Dunbar, 1988). When those shapes were presented in colours that conflicted with their newly learned names, shape naming exhibited the attributes of a controlled process (it was slower, and subject to interference from the colour in which the shape was displayed), whereas colour naming appeared to be the automatic process (faster, and unaffected by the shape's name). At the same time, findings were reported suggesting word reading, a process considered to be canonically automatic (LaBerge & Samuels, 1974; Posner & Snyder, 1975) could be shown to rely on attention and/or control (Kahneman & Henik, 1981). Such observations presented a paradox for the assumption that a process was either controlled or automatic; rather, it seemed to depend on the context in which the process was executed.

Learning. A closely related observation was that controlled processes could become automatic with practice. In particular, in one of the most elegant set of studies in cognitive psychology, Schneider and Shiffrin (1977) showed that if a task was practised extensively, it developed all of the signs of automaticity: It became faster, less effortful, and less subject to control. Critically, however, this required that the association between stimuli and responses remained fixed. If these varied, automaticity did not develop even over the course of thousands of trials with the same stimuli and responses. Interestingly, recent evidence suggests that, under some circumstances, processes with the signature of automaticity can develop much more quickly (Meiran, Pereg, Kessler, Cole, & Braver, 2015—see below). The extent—and type—of training required for a process to become automatic (i.e., less reliant on control) remains a critical area of inquiry, and may hold important clues to the mechanisms involved.

Continuum of automaticity and control. Taken together, the observations reviewed above have been interpreted as evidence that controlled and automatic processing define the ends of a continuum, and that the place occupied by a given process along the continuum is a function of both learning (i.e., the number of times that exact process has been executed) and the context in which it occurs (i.e., what other processes are engaged at that same time). This characterisation has been formalised in a variety of models, ranging from symbolic production system models (e.g., Anderson, 1983) to neural network models (e.g., Cohen et al., 1990; Cohen, Servan-Schreiber, & McClelland, 1992). The latter directly address the

graded nature of automaticity, attributing this to the strength of the processing pathway required to perform a task (which is directly determined by learning), relative to the strength of pathways supporting any processes with which it must compete. In such models, control augments the sensitivity of a pathway to its inputs, allowing it to compete more effectively with other pathways that carry interfering information.

Flexibility

A hallmark of the human capacity for cognitive control is its flexibility: the remarkable ability to rapidly configure and execute a seemingly limitless variety of behaviours, including ones that have never before been performed (e.g., Meiran et al., 2015). Most current theories of cognitive control assume that this relies on the activation of control representations that serve as internal context, guiding processing in the parts of the system required to implement goal-relevant processes or behaviour. In symbolic architectures (e.g., ones that use production system architectures, such as Anderson, 1983), this is assumed to rely on representations in declarative memory (e.g., goal representations). In neural models, it is generally assumed to rely on the activation of context representations in the prefrontal cortex (e.g., Miller & Cohen, 2001). Symbolic models readily afford flexibility, as they have at their core the ability to bind control representations (e.g., variables) to arbitrary values, and compose these in arbitrary ways to implement new behaviours. However, whether arbitrary variable binding and full compositionality are implemented in the brain is less clear. At the least, humans exhibit limitations in abilities that are trivial to implement in truly symbolic systems, such as multi-digit mental arithmetic or parsing recursively embedded phrases (e.g., the mouse the cat the dog chased scared squealed). Understanding the flexibility of human behaviour in terms of mechanisms that approximate symbol processing, or that use altogether different computational mechanisms, is one of the major challenges for the study of cognitive control (O'Reilly et al., 2013). Efforts to address this challenge have brought into focus four functional requirements that such a system must satisfy: (a) a *representational code* sufficient to span the seemingly limitless range of control-dependent behaviours; (b) the ability to *acquire* (or configure) such a code from experience; (c) the ability to *update* the currently active control representation(s) in a context-appropriate manner; and (d) the ability *select* the appropriate representation to activate. Understanding how these requirements are met has become an important focus of research on cognitive control.

What code do control representations use? This is perhaps the most fundamental question that confronts research on cognitive control: What is the form of the representations used to flexibly guide performance? Some have argued against an explicit code, suggesting that a sufficiently large, randomly connected network may be powerful enough to explain the flexibility of control-dependent behaviour (Rigotti et al., 2013; Susillo, 2014). However, this begs the question of how, out of the vast number of possibilities, the system can be configured immediately (e.g., under instruction) to implement the precise combination of processes needed to implement an arbitrary, novel task. Others have proposed that, as in fully symbolic systems, there must be some more systematic, combinatorial code—a 'vocabulary'—that efficiently spans the space of possibilities (e.g., Eliasmith et al., 2012; Plate, 1995). Whether random or systematic, characterising the nature of the representations on which control relies remains one of the greatest challenges of cognitive neuroscience. Meeting this challenge may rely on progress in addressing another, closely related challenge: understanding how control representations arise through learning and development.

How are control representations acquired? The capacity for cognitive control is clearly one that emerges over the course of development, almost certainly under the regulation of genetic factors, but equally clearly shaped by experience. Understanding this developmental

process has been the focus of considerable empirical study (Diamond, 2013; Hanania & Smith, 2010; Munakata, Snyder, & Chatham, 2012; Chapter 26 by Cohen & Casey in this volume). There are at least two lines of theoretical work that address the learning mechanisms underlying cognitive control. One has explored the idea that, with the appropriate neural architecture, simple reinforcement learning algorithms can extract representations from experience that are sufficiently abstract and compositional as to approximate symbol-like processing, even if they are not fully symbolic in the sense of supporting arbitrary variable binding (Kriete, Noelle, Cohen, & O'Reilly, 2013; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). Another line of work has explored an extension of simple reinforcement learning—referred to as 'hierarchical reinforcement learning'—that is sensitive to the nested goal–subgoal structure of many tasks, and exploits this to extract representations that can be used to control behaviour over a wide range of tasks (e.g., Botvinick, Niv, & Barto, 2009; Frank & Badre, 2012). These approaches complement one another, and an integration of the insights gained from each promises to advance our understanding of the code used to represent control signals, and how it emerges with experience. This work also points to a third requirement for flexibility of behaviour: the ability to update control representations in a context-appropriate manner.

How and when are control representations updated? All model architectures that address cognitive control implement some mechanism for regulating the activation and maintenance of control representations. In production systems, this is typically managed by the firing of productions that are responsible for updating the contents of working memory. However, several critical features vary across models: How many productions can fire at once; whether they do so synchronously (all updates occur at once) or asynchronously (whenever a relevant production fires); and the 'conflict resolution' mechanisms required to select which of a set of competing productions are allowed to fire at a given time. Similar issues arise in neural network architectures, for example, whether to update representations continuously (as in simple recurrent networks; e.g., Botvinick & Plaut, 2004; Cleermans & McClelland, 1991; Elman, 1990) or at discrete intervals. The latter is usually implemented using a gating mechanism that regulates access to parts of the system responsible for representing and maintaining control signals. Such a mechanism has been implemented both in models used to simulate human performance and brain function (Braver & Cohen, 2000; Chatham & Badre, 2015; Frank, Loughry, & O'Reilly, 2001; Todd, Niv, & Cohen, 2008; Zipser, 1991), as well as in machine learning applications (Hochreiter & Schmidhuber, 1997; LeCun, Bengio, & Hinton, 2015). Such models, applied to human function, make useful contact with the neural mechanisms involved. However, an important challenge for these models is to make contact with the rich set of experimental findings concerning human performance in domains such as task switching that address the dynamics of updating of control (e.g., Collins & Frank, 2013; Gilbert & Shallice, 2002; Reynolds, Braver, Brown, & Van der Stigchel, 2006).

One issue that has come to the fore, concerning the timing of control updating, is whether this happens in anticipation of the need to change the control representation or 'just-in-time' (that is, when the first indication occurs of the need for a new control state). These strategies have been referred to as 'proactive' and 'reactive' control, respectively (Braver, 2012; Chapter 9 by Chiew & Braver in this volume). This issue raises interesting questions about how control can be optimised, as well as its interaction with other memory systems—issues that will be touched on further below.

Another critical question is how the system *learns* when to update control representations. One intriguing answer to this question, suggested by neurophysiological data, is that the gating signal may be tightly coupled with, or may even rely on the same mechanisms used for reinforcement learning. The latter respond to reward prediction errors, which are elicited by

a stimulus that itself was not predicted, but that predicts a later reward (Montague, Dayan, & Sejnowski, 1996). This is precisely the timing required for a gating signal: an event that, itself unpredicted, signals the possibility of greater reward if control is redeployed. Thus, coupling the gating signal to the reinforcement learning signal may allow the system to discover when the gating signal should occur (Braver & Cohen, 2000; Frank et al., 2001). Recent work has suggested that gating of the output of the control system may be as important as gating its input (e.g., Chatham, Frank, & Badre, 2014; Kriete et al., 2013). These ideas offer a rich avenue for future research on how the control system develops and how, when it goes awry, it may manifest in clinical conditions.

How are control representations selected for activation? The adaptive value of a control system relies critically on its ability to determine not only *when* control signals should be updated, but specifically *which one(s)* to engage at a particular time. The conflict resolution mechanisms in symbolic models partially address this point, but typically do so on the basis of the structural properties of the opportunities (e.g., which properties are most highly specified) rather than on motivational factors, such as effort and reward. Recent efforts have begun to focus on the latter, suggesting that control signals are selected based on a cost-benefit analysis that takes account of the expected value of the control signals in contention, and selects the one(s) that maximise this value (Shenhav, Botvinick, & Cohen, 2013; see also Chapter 10 by Kool et al. in this volume). This kind of normative approach is a promising avenue for research on control, as discussed in the final section of this introduction.

An intimately related question is how the value of prior experience (i.e., knowledge already accrued about expected outcomes for control signals) is balanced against the value of gaining new experience—a tension that is commonly referred to as the explore–exploit trade-off. Managing this trade-off successfully is a fundamental requirement for any agent that can adapt in non-stationary environments (e.g., Cohen, McClure, & Yu, 2007; Gittins & Jones, 1974; Kaelbling, Littman, & Moore, 1996; Krebs, Kacelnik, & Taylor, 1978; Pratt & Sumpter, 2006; Watkinson et al., 2005). Although there is no fully general solution to the explore–exploit trade-off, recent work has begun to examine how humans manage this problem (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Wilson, Geana, White, Ludvig, & Cohen, in press). These lines of work examine how, in choosing a course of behaviour, agents should weigh the value of acquiring information (i.e., exploration, in the service of learning, and thus as a proxy for future reward) against the value of immediate reward (i.e., exploitation). Here, as with updating mechanisms, there are proposals about the role that neuromodulatory mechanisms may play in regulating this critical function of control (Aston-Jones & Cohen, 2005; Yu & Dayan, 2005) that may help index this function, and provide clues to both its normal operation and the role it plays in clinical conditions. A consideration of the explore–exploit trade-off also suggests an additional point of contact between control and motivational constructs, such as boredom. Recent findings have begun to suggest that, just as effort may reflect a phenomenological correlate of the opportunity cost associated with performance of a given task with respect to *reward*, so boredom may reflect the opportunity cost with respect to *information*—that is, boredom may signal the value of exploration (Geana, Wilson, Daw, & Cohen, under review). These lines of work are beginning to sketch the outline of a more formal, comprehensive, and normative model of control.

Unitary or Multiple Constructs

The earliest theories of cognitive control were explicit in suggesting that it relied on a single, central processing mechanism. This was based in large measure on the observation of capacity constraints (e.g., dual-task interference) and the assumption that this reflects the limited capacity of a central mechanism. As discussed above, this claim has been challenged by the

suggestion that controlled processing may lie along a continuum, and that interference effects reflect interactions among the local processes that control is engaged to regulate, rather than an intrinsic limitation of a centralised mechanism. Nevertheless, even if control can be engaged along a continuum and its capacity is not limited, it is meaningful to ask whether control relies on a unitary mechanism, or on disparate domain-specific mechanisms. This question closely parallels ones that have been raised with respect to other constructs, such as intelligence, working memory, and attention. Here, as in those cases, an intermediate answer seems most likely: that control relies on mechanisms of a particular *kind,* implemented in a wide range of domains, with different parameterisations that reflect specific features of those domains, but that share common fundamental attributes. If this is so, then understanding this 'family resemblance' may help guide modelling, hypothesis generation, and empirical inquiry. Thus, it may be most useful to ask: What are the fundamental functions and attributes that define cognitive control, and what are the kinds of mechanisms that can implement these functions and attributes?

The sections above—and many of the chapters of this book—are meant to address these questions. In summary, the following seem to be a core set of functions and attributes of cognitive control: (a) the ability to support control representations as a form of internal context, that serve as signals to guide processing elsewhere in the system; (b) the use of such signals to avert interference that can arise from cross-talk among processes sharing overlapping pathways; (c) constraints on the capacity for control-dependent processing that may reflect more about the propensity for such cross-talk, and the need to limit it, than limitations in the control system *itself;* (d) the ability to update control representations in a manner that is sensitive to the circumstances in which this is needed; and also to (e) the opportunity costs that constraints on control-dependent processing impose, weighed against the opportunities that controlled processes afford for both expected reward (exploitation) and/or to gather useful information (exploration and learning). Progress is being made in understanding the neural architecture and mechanisms underlying these functions and attributes, which provide additional support for the supposition that, at the least, the construct of cognitive control remains a coherent and useful one.

Relationship to Other Psychological Constructs

From its inception, the construct of cognitive control has been inextricably intertwined with several others in the psychological literature, and thus it seems important to consider how these are related. Several of these stand out: executive function, intelligence, volition, attention, working memory, and inhibition. In addition, recent research has begun to focus on the role of cognitive control in two other domains of function: self-control and long-term memory.

Executive Function and Intelligence

Executive function. This term has a long history in the psychological literature (e.g., Bianchi, 1895; Luria, 1966), extending into early work in cognitive psychology (e.g., Baddeley & Hitch, 1974; Shallice, 1982). It seems impossible to distinguish the use of 'executive function' from the construct of cognitive control and, at least within the cognitive psychological literature, 'cognitive control' has largely replaced the use of 'executive function'. This is likely for historical reasons (e.g., the close association of 'executive function' with older neuropsychological constructs and particular batteries of tasks). In any event, it seems reasonable to treat these as synonyms.

Intelligence. Since Spearman introduced the construct of 'general intelligence' (*g;* Spearman, 1904), it has been both captivating and controversial, and inextricably intertwined with the concept of executive function. Like executive function, it has also been associated consistently with frontal lobe function (e.g., Duncan et al., 2000). The question of whether intelligence should be treated as a unitary construct (as proposed by Spearman [1904]), or reflects the operation of a collection of domain-specific faculties parallels, in many ways, a similar question about cognitive control. Answers to these questions may bear on a deeper understanding of the relationship between them, and in particular, whether intelligence reflects the functioning of mechanisms beyond those responsible for cognitive control. What does seem clear, however, is that the capacity for cognitive control represents a necessary, if not sufficient, condition for intelligence (e.g., Kane & Engle, 2002).

Volition

The earliest treatments of cognitive control proposed that the mechanisms involved were responsible for conscious, voluntary action, and intention. This had strong intuitive appeal, and continues to be an interesting and potentially valuable area of inquiry (Dehaene & Naccache, 2001; Graziano & Kastner, 2011). In part, this is because it seeks to satisfy our intrinsic curiosity about the compelling nature of our experience as 'volitional' agents, and because it may lend insights into fundamentally important social and moral questions (such as responsibility for action; e.g., Bratman, 1987; Vargas, 2013). Progress has been made in identifying the correlates of conscious states (e.g., Cleermans, 2007; Dehaene & Changeux, 2011; Schurger, Pereira, Treisman, & Cohen, 2010) and volition (Haggard, 2008; Soon, Brass, Heinze, & Haynes, 2008). However, the extent to which these are co-extensive with those that engage control remains an open and interesting question.

Attention

The construct of cognitive control grew out of the literature on attention, and remains intimately bound to it. One might even argue that these constructs are inseparable, with attention referring to one of the most fundamental functions of control: the selection of some processes for engagement over others (e.g., Ardid, Wang, & Compte, 2007; Cohen et al., 1990; Deco & Rolls, 2005). However, the question of whether 'attention' should simply be considered as a function of control—and nothing more—raises some of the questions discussed above concerning the scope of the construct of control. For example, is it meaningful to assume that the mechanisms responsible for sensory selection (to which the word *attention* is most commonly applied) reflect the operation of cognitive control? Do they share features in common with those responsible for the selection of actions? Similarly, does the 'exogenous' and seemingly automatic engagement of attention (e.g., the 'capture of attention' by salient events, such as a loud sound) engage mechanisms of cognitive control, and are those the same as or in some meaningful way similar to those responsible for 'endogenous' (or 'strategic') forms of attention (e.g., responding to a verbal instruction)? At the least, it has been known for quite some time that these exhibit different dynamics (e.g., Neely, 1977; Petersen & Posner, 2012; Posner & Cohen, 1984). Once again, an answer to these questions, and whether it makes sense to consider attention to be a function of control in all circumstances and at all levels of analysis will only be answered (and may ultimately be rendered irrelevant) with a deeper understanding of the mechanisms involved.

Working Memory

The construct of working memory, first formulated in the context of theories about executive function (Baddeley & Hitch, 1974), referred to a buffer that stored information required for executive processes to operate (e.g., the intermediate products of a computation). Symbolic models proposed that working memory (defined as the activated state of information in long-term memory) served not only to buffer intermediate products of computation, but also goal representations used to guide behaviour (Anderson, 1983). Neural network models of control have emphasised the latter (Cohen et al., 1990; Dehaene & Changeux, 1991; O'Reilly, 2006). In general, there seems to be a consensus that cognitive control relies critically on at least one component of working memory: the activation and maintenance of control representations used to guide processing (Miller & Cohen) and, furthermore, that a critical function of control is to regulate the contents of this component of working memory (O'Reilly, 2006). However, as with attention, it is not yet clear how the boundaries of this construct coincide with those of cognitive control (see also Chapter 3 by Meier & Kane in this volume). This is complicated by overlapping use of the terms 'working memory' and 'short-term memory', both of which refer to the maintenance of information in an activated state. Like the relationship between cognitive control and attention, clarity along these lines is most likely to come from a more detailed understanding of the underlying mechanisms, and their relationship to those involved in attention and control (e.g., Ikkai & Curtis, 2011), rather than from further attempts to refine these definitions in the absence of such an understanding.

Inhibition

Directed versus competitive inhibition. The associations of inhibition with executive function and the frontal lobes dates back to the report about Phineas Gage (Harlow & Martyn, 1868), and neurological studies in the beginning of the last century (Adie & Critchley, 1927; Brain & Curran, 1932; Marinesco & Radovici, 1920) identifying reflexes in infants that disappear in adults, but reappear in patients with damage to the frontal lobes. These 'frontal release signs' continue to be used in clinical practice to identify frontal lobe damage. This phenomenon, together with the observation of primitive (e.g., 'utilisation') behaviours in monkeys and humans with frontal lobe lesions, led to the belief that a cardinal function of the frontal cortex (and the executive function that it supports) is the inhibition of reflexive and/or habitual behaviours (e.g., Bianchi, 1895; Fuster, 1980; Lhermitte, 1983). This belief persists, in many quarters, in the common and steadfast assumption that frontal control mechanisms provide direct and specific inhibition of automatic (e.g., habitual) processes (Buckholtz, 2015). However, there is little direct support for this assumption beyond the domain of neurological reflexes. An alternative to directed inhibition is competitive inhibition: by facilitating selected processes, control allows those processes to compete more effectively with interfering processes (e.g., by way of lateral inhibition, or other 'normalising' mechanisms). It may be difficult, and even impossible, to distinguish between directed and competitive inhibition on theoretical grounds alone. However, competitive inhibition seems both more parsimonious (it should be easier to select and support a single process than direct inhibition towards all potential competitors), and it is consistent with general principles of neural organisation (e.g., the scarcity of long-range inhibitory projections, and the abundance of lateral inhibitory interneurons). Nevertheless, an adjudication of these possibilities, or the identification of an intermediate solution, is likely to require detailed neurophysiological investigation.

Global inhibition and 'stopping'. While the mechanisms of inhibition underlying selection remain to be identified, it has become increasingly clear that the brain implements, and the control system has access to a 'stopping mechanism' that effects a more general or global form of inhibition. This is suggested both by behavioural evidence (e.g., using stop-signal and go-no-go tasks; Logan, 1994; Verbruggen & Logan, 2008), and by neural evidence concerning the basal ganglia, and in particular the subthalamic nucleus (e.g., Aron & Poldrack, 2006; Frank, 2006). One hypothesis is that such systems provide a 'brake' that may reflect an asymmetry in both mechanism and outcome between rewarding and perilous outcomes: In general, the risks associated with erroneous action may be greater or more frequent than those associated with inaction. Such a stopping mechanism may also serve a more refined function in the control of behaviour, by setting response thresholds used to regulate speed–accuracy trade-offs in the service of optimising reward (e.g., Bogacz et al., 2006; Cavanagh et al., 2011; Gold & Shadlen, 2002). Understanding the mechanisms by which control overrides inappropriate responses—whether by directed, competitive, or some more global form of inhibition—remains an important goal not only for basic research, but also for understanding failures of control and, in particular, failures of self-control that are a prevailing social and clinical concern.

Self-Control

The study of self-control has become one of increasing importance, at both the individual and social level (see Chapter 25 by Davisson & Hoyle in this volume). At the individual level, it is obvious that failures of self-control are a fundamental feature of a wide range of clinical disorders, from obsessive-compulsive disorder to drug addiction and gambling (Chapter 32 by Chaarani et al. in this volume). It is also becoming increasingly apparent that problems of self-control are responsible for dysfunctional behaviours in otherwise healthy individuals (such as failures to save adequately for retirement) and at the societal level (e.g., energy use policies). Although, like many of the constructs discussed above, there is no scientifically accepted definition of the term *self-control*, in common use it refers to 'the ability to control oneself, in particular one's emotions and desires'. Once again, it remains to be determined whether the construct of cognitive control, as traditionally defined and studied in the experimental laboratory, extends seamlessly into the domain of self-control in the real world (e.g., Buckholtz, 2015). There are at least two ways in which self-control may involve domain-specific factors: (a) the nature of the process over which control presides; and (b) what appears to be the fundamentally intertemporal nature of decisions involving self-control.

Value-based decision making. If a primary function of control is to choose behaviours that maximise value (see the section above titled 'Effort and Motivation'), then it seems natural to include decisions driven by emotions and desires in its scope (e.g., see Chapter 23 by Winecoff & Huettel in this volume). These can be considered as a class of automatic processes that pose challenges to controlled processing similar to more 'cognitive' ones (e.g., word reading in the Stroop task). Indeed, there is a large literature that exploits this correspondence to infer the influence of emotional and valuation processes on behaviour (e.g., Fazio & Olson, 2003; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Chapter 24 by Krebs & Woldorff and Chapter 22 by Pessoa in this volume), suggesting that the role of control in regulating cognitive processes may share much in common with mechanisms involved in regulating emotional ones (e.g., Gross & Thompson, 2007). At the same time, affective processes may exhibit dynamics that distinguish them from other types of automatic processes. For example, the 'force' of a desire may increase with time, and the control of such processes often exhibits a stereotyped recovery-and-relapse cycle (e.g., as observed in addiction, dieting, etc.) that is not commonly observed in other domains. These factors pose both theoretical and experimental challenges that may be unique to the domain of self-control.

Intertemporal choice. Decisions about self-control seem invariably to involve choices between one option that is immediately desirable (e.g., a piece of cake on the table) and another that carries greater future value (dieting), a type of decision often referred to as intertemporal choice. Thus, self-control may fundamentally involve not only inhibition (as discussed above), but also intertemporal choice. This is consistent with a growing literature suggesting that intertemporal choice involves a competition between automatic and controlled processes (Figner et al., 2010; Kahneman, 2003; McClure, Laibson, Loewenstein, & Cohen, 2004). Indeed, one intriguing conjecture is that intertemporal choice is fundamental not only to *self*-control, but to control in general. In other words, the immediacy of the reward associated with an outcome may be a critical feature in determining which of two (or more) competing processes demands control. In the Stroop task, the greater ease of reading the word might be viewed as a form of immediate reward, whereas the benefits of overriding this response in order to name the colour come when performance is rewarded, which is usually later. From this perspective, the demands for control might be viewed as intimately bound to intertemporal choice, and control mechanisms as the steward of our future selves. Nevertheless, intertemporal choice involving affective processes and self-control may impose distinct requirements (e.g., due to the greater salience of rewards, and/or greater temporal asymmetries) and may therefore demand distinctive mechanisms of control.

Long-Term Memory, Prospective Memory and Planning

Memory search. Mechanisms that actively represent internal context have come to occupy a central role in theories about the encoding and retrieval of information from long-term memory. Tulving's context-encoding hypothesis proposed that features of the context in which a memory was formed are encoded along with the memory itself, and that retrieval involves a form of 'mental time travel' that reinstates the context in which the memory was formed to facilitate retrieval of the memory (Tulving, 2002). The temporal context model (TCM, Howard & Kahana, 2002; Polyn, Norman, & Kahana, 2009) extends this idea, suggesting that internal, actively maintained representations—such as goals and intentions—are a particularly useful form of context that can identify the time at which a particular memory was encoded. On this view, retrieval involves reinstating the context representation active at the time of memory encoding and, by association, the memory itself. This beautifully ties together the role of context representations in control and long-term memory, and the foundations for a mechanistic understanding of the processes involved.

Prospective memory and planning. Like most studies of cognitive control, TCM focuses on its regulative functions—the selection of a process (in this case, memory retrieval) for current execution. However, an important and growing area of research is on the interaction between control and long-term memory in the service of prospective memory and planning. Imagine the following example: I ask you to perform the colour naming task in 2 s. It is almost certain that you will engage the required task representation(s) immediately. However, if instead I ask you to do it when I return to the room in an hour, it is just as certain that you will *not* engage and maintain those representations while I am gone. Rather, you will do so when I return. Critically, it is likely you will be able to do this without my having to instruct you when I return. This represents a form of prospective memory ('remembering' to do something in the future), and a simple form of planning.

There is increasing evidence that this ability to program a controlled process for future execution relies on an interaction between control mechanisms and episodic memory (e.g., Cohen & O'Reilly, 1996; Einstein & McDaniel, 2005; Gollwitzer, 1996). On this account, when an instruction is presented (or a plan is conceived), an association is formed in episodic memory between the control representation required to execute the necessary behaviour and

the future conditions under which it should be executed (e.g., my reappearance in the room). Then, when those conditions occur, the association in episodic memory elicits retrieval of the control representation, which is re-engaged (e.g., gated) in working memory, and the task is executed. Although this idea that control representations can be 'cached' in episodic memory is compelling, it poses many new questions. For example, when is this mechanism used rather than the immediate activation of the control representation (a question closely related to the use of proactive versus reactive forms of control discussed in the previous section titled 'Flexibility'; e.g., Bugg, McDaniel, & Einstein, 2013; Meiran et al., 2015), and what are the factors that influence this decision? When choosing to defer, what features of the future state are chosen to associate with the control representation? To what extent can the association between control representation and episodic memories emerge passively, through experience (cf. Chapter 4 by Egner in this volume)? Answers to these questions promise to unravel one of the greatest mysteries of the human brain: how it supports the ability to plan for the future.

Theoretical Considerations

Normative Theory

The dominant approach to research on cognitive control has been to characterise the properties of control-dependent processing, use these to infer candidate mechanisms, and design experiments to test those mechanisms. Although this has generated considerable progress, there is another, complementary approach that has been conspicuously scarce in research on cognitive control: the construction of normative theory. This is sometimes referred to as rational analysis, or the ideal observer method (e.g., Barlow, 1981; Tanner & Swets, 1954). This seeks to identify the optimal computation for a function of interest, which is then used to generate hypotheses about the mechanisms involved. Although it is rare (though not unprecedented) that natural systems implement fully optimal mechanisms, this approach provides a rational guide for generating hypotheses, and a formally rigorous framework within which to test them. It has driven considerable progress in many domains of science, including psychology (Anderson, 1990; Geisler, 2003; Tenenbaum, Griffiths, & Kemp, 2006). The scarcity of this approach in research on cognitive control is particularly surprising, given that the *optimisation* of behaviour can be viewed as the fundamental purpose of control. This is the definition of control used by systems theory in engineering and, as noted above, inspired the earliest thinking about control in the context of human behaviour (e.g., Wiener, 1948; Miller et al., 1960).

A critical step in normative theory is defining the 'objective function' being optimised; that is, in the context of cognitive control, the goal that the behaviour is intended to achieve. Identifying the objective function for cognitive control poses a serious challenge, given the broad scope of processes, behaviours, and goals it can serve. Once again, this raises questions about the extent to which 'cognitive control' should be treated as a unitary construct, a class of mechanisms sharing a family resemblance, or a disparate set of domain-specific mechanisms. Nevertheless, several lines of work have begun to take on the challenge of this approach.

Task-level optimisation. One approach has been to focus on the role of control in optimising a particular task. An example of this is work on two-alternative forced choice decision-making tasks. Dramatic progress has been made in identifying and characterising the mechanisms involved in such simple decisions, at both the psychological and neural levels of analysis (e.g., Gold & Shadlen, 2007; Ratcliff & McKoon, 2009), and in conducting normative analyses of performance (e.g., Bogacz et al., 2006). This has provided a formally rigorous

framework within which to interpret psychological constructs (such as expectations, attention, and the speed–accuracy trade-off), and for quantifying performance associated with optimal control against which human performance can be compared (e.g., Simen et al., 2009). This work serves as an example of how normative theory can be used to study the engagement of cognitive control in a given task domain, of which other examples are now beginning to emerge (e.g., Verguts, Vassena, & Silvetti, 2015; Wiecki & Frank, 2013; Yu, Dayan, & Cohen, 2009).

Meta-level optimisation. A complementary approach has been to consider the problem at the broadest level: How does the control system determine which tasks or goals should be pursued at a given time? One approach is to ascribe this to learning (e.g., Verguts & Notebaert, 2008); another is to consider it as an optimisation problem of its own, nesting the optimisation control for a given task within a higher-level optimisation of the choice among tasks. An example of this is the expected value of control (EVC) theory (Shenhav et al., 2013). This assumes that investment in a control-demanding task improves the probability of reward, but that this carries a cost that scales with the amount of control invested. The EVC theory proposes that control is allocated across tasks based on a cost-benefit analysis of this trade-off, so as to maximise the overall rate of reward. This provides a rigorous framework within which to analyse, and make predictions about the allocation of control in a given task environment.

Optimisation under constraints: Bounded rationality. The cost of control must be taken into account by any normative theory of control. Understanding this cost—its functional form, and whether and how it varies across individuals and domains of behaviour—is an important direction for research. As discussed above, it seems likely that a central factor is the capacity constraint on control, which imposes an opportunity cost: Investing control in one task forgoes the opportunities for reward afforded by others. This reinforces the importance of understanding the nature and source of the constraints on control. Here, computational analyses are proving useful, as discussed further below. Consideration of how control is optimised in the face of a budget represents an instance of an approach to normative theory, broadly referred to as 'bounded rationality' (Simon, 1955, 1992), that has begun to attract growing attention (Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, Lieder, & Goodman, 2015; Howes, Lewis, & Vera, 2009). This assumes that optimisation must take account not only of the system's objectives, but also the constraints under which it must operate—akin to tuning a radio as best as possible to a weak station. One criticism of this approach is that it is possible to explain any pattern of performance post hoc, by conjuring a set of constraints under which the observed performance would be optimal. However, rather than a problem, this can be viewed as a valuable step in the scientific process, the next step of which is to cast those constraints as hypotheses and use them to generate new predictions about performance. This approach has begun to show potential (e.g., Balci et al., 2011; Lieder & Griffiths, 2015), and is a promising avenue of research for the study of control.

The statistics of natural tasks. The properties of the control system itself, the processes over which it presides, and the neural architecture in which these are implemented represent important sources of constraint (Botvinick & Cohen, 2014). However, an equally important one is the *environment* in which the system operates. It is, after all, the environment to which an adaptive system adapts. Therefore, taking account of the structure of the environment should provide important clues about the structure of systems responsive to it. For example, recent progress in understanding the function of the human visual system has been driven in large measure by a characterisation of the statistics of natural scenes that have shaped its architecture (Simoncelli & Olshausen, 2001). The study of cognitive control stands to benefit from a similar approach. The human brain is clearly better at carrying out some kinds of tasks (e.g., crossing a busy road) than others (long division), likely because it has evolved to solve

a wide, but not random, set of tasks required to survive in the natural world. Characterising the structure and statistics of tasks should be an important priority for research on cognitive control.

Computational Trade-Offs in Representation and Processing

The constraints on processing considered by the approaches discussed above have largely been physical (e.g., limits in the amount or type of data available to the system).

However, a more fundamental set of constraints may be related to trade-offs inherent to all computational systems. One such trade-off, which has been central to the construct of cognitive control since its inception, is between serial and parallel processing. A second trade-off that has begun to attract attention is between model-free and model-based processing.

Serial versus parallel processing. This distinction was central to the original formulations of controlled and automatic processing (e.g., Shiffrin & Schneider, 1977). Reliance on a central, limited-capacity mechanism was assumed to impose a serial constraint on controlled processing, akin to the sequential execution of a programme by the central processing unit of a standard computer (i.e., one that implements a von Neumann architecture). Conversely, it was assumed that automatic processes can be executed in parallel (i.e., simultaneously without penalty), akin to the 'embarrassing' parallelism common in multi-node computer clusters (that is, running many unrelated jobs on different nodes simultaneously). The serial constraint on controlled processing has also been a central assumption of two of the most fully developed and influential models of cognition (ACT-R and SOAR; Anderson, 1983; Newell, 1990). However, the necessity of this assumption has been challenged by models using similar architectures that weaken or eliminate the serial constraint on controlled processing (e.g., Meyer & Kieras, 1997; Salvucci & Taatgen, 2008). The debate about whether there is a 'central bottleneck' in cognitive control has also been subject to intense empirical inquiry, centred largely around the observation of the psychological refractory period (Pashler, 1984)—a delay in performance associated with the attempt to perform two or more tasks at once, interpreted as evidence that they are being queued for serial execution. The interpretation of this finding in terms of a central bottleneck has been challenged (e.g., Howes et al., 2009; Schumacher et al., 2001); however, several theories continue to assume that control relies on a centralised mechanism (Duncan & Owen, 2000; Roca et al., 2011; Tombu et al., 2011).

The debate about whether controlled processing relies on a central, serially constrained mechanism highlights, and is complicated by, another problem: In practice, it may be very difficult to distinguish between rapid serial and truly (concurrent) parallel processing (e.g., Townsend, 1972). The ability to do so, and the relevance of doing so, depends on the temporal resolution—both of the measurement, and of the outcome of interest. For example, the serial updating of pixels on a computer display could be detected by an oscilloscope with a temporal resolution of greater than 100 Hz, but not the human visual system with a temporal resolution of less than 60 Hz (which perceives the update as synchronous). This suggests that a definitive answer to the question of whether controlled processing is purely sequential, or can support parallel processes, will require, like other questions, finer-grained measurements and possibly neurobiological evidence.

That said, there is another standpoint from which to view the distinction between serial and parallel processing, and its relationship to cognitive control, that may lend clarity and coherence to the array of phenomena associated with controlled processing. This can best be appreciated by considering another kind of parallelism, captured by parallel distributed processing (PDP) or 'connectionist' architectures (Rumelhart, McClelland, & the PDP Research Group, 1986). Here, rather than the number of behavioural tasks that can be performed at once, the appeal of parallelism is the number of constraints that can be taken into account

in computing the solution to a single problem—a process sometimes referred to as 'multiple constraint satisfaction'. Each individual unit in a PDP network can be thought of as representing a constraint (or 'micro-process'), and their interaction serves to take account of the mutual influence that these have on each other in parallel, averting the costs of a combinatorial explosion that would be incurred by doing so in serial. Mutual constraint satisfaction is a hallmark of functions at which the human brain excels (such as face perception and natural language processing). PDP models have been used to understand how the human brain computes these functions, and recent advances in machine learning have begun to approximate these abilities using artificial neural networks (LeCun et al., 2015).

The success of PDP networks lies in the extensive interactions among individual processing units—sometimes referred to as fine-grained parallelism. This is in contrast to coarse-grained (or 'embarrassing') parallelism that supports the execution of multiple independent processes at once. Not surprisingly, there is a trade-off between these types of parallelism: The extent to which a network supports fine-grained interactions among its units in performing a task (in the service of mutual constraint satisfaction) and shares the representations involved across multiple tasks (supporting generalisation) is in tension with the extent to which it can support the performance of multiple such tasks at once (Musslick et al., 2016). This suggests there may be a fundamental link between the richness of interactions among the processes involved in performing a task (e.g., recognising a face), and the imposition of a serial constraint on that performance (e.g., finding a face in a crowd). This relationship may also help explain the canonical trajectory during learning from control-dependent to automatic processing (e.g., Schneider & Shiffrin, 1977), in terms of a transition from interactive, generalisable representations that rely on fine-grained parallel processing (and thus demand seriality), to independent, dedicated representations that afford coarse-grained parallel execution (i.e., multitasking). A better understanding of this relationship between representation and the trade-off between serial versus parallel processing may not only offer a new way to frame important, long-standing questions about control, but also new approaches to measurement (e.g., Musslick et al., 2016) and, potentially, intervention.

Model-based versus model-free processing. This distinction, with origins in the work of Tolman (1948), has recently been cast in terms of formal learning algorithms and regained the attention of psychologists and neuroscientists (e.g., Dickinson & Balleine, 2002; Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011; Chapter 11 by de Wit in this volume). In a model-free system, actions are selected based on direct associations from the stimulus to the response, learned through trial and error, and without a representation of potentially mediating factors (Sutton, 1988). In a model-based system, actions can also be evaluated by computing and evaluating *potential* courses of action, using an 'internal model' of the possibilities that can include intervening states (i.e., between the stimulus and response). Although the former is more efficient at processing (i.e., it requires less computation, and thus can respond more quickly in a *given* environment), the latter is more flexible (it can adjust more quickly to *changes* in the environment, by modifying the model rather than relying exclusively on trial-and-error-experience).[2] The construct of model-based processing has played a central role in theories about the role of 'internal replay' in learning and decision making (e.g., Sutton, 1990), in which rehearsal of past sequences of occurrences is used as a proxy for actual experience in learning, and in planning future actions (Gershman, Markman, & Otto, 2014; Redish, 2016; Shohamy & Daw, 2015). Recently, it has been proposed that the distinction between model-based and model-free processing may reflect the differential engagement of controlled and automatic processing—with model-based processing relying on control mechanisms, whereas model-free reflects the operation of more automatic ones—and empirical evidence has begun to accrue in support of this (Deserno et al., 2015; Otto, Skatova,

Madlon-Kay, & Daw, 2015). This offers a rich theoretical framework within which to explore the role of controlled processing in learning and memory, including its role in prospective memory and planning as discussed above.

Bringing the lines of work described above—on the relationship of representation to parallelism, and on model-based versus model-free processing—into contact with one another, and using them as a framework for building models of cognitive control is a particularly promising direction for future theoretical work.

Summary

The construct of cognitive control is a foundational one in cognitive psychology. The phenomenology that initially inspired the construct is compelling—in particular, its association with 'mental effort', the manifest constraints on its capacity (in both number and duration), its apparent sequentiality, and its place in the trajectory of learning—and, for the most part, these are empirically validated. However, defining cognitive control in a more rigorous way and identifying the mechanisms that govern its operation have been a challenge. Recent work at the intersection of cognitive psychology, neuroscience, and computer science has begun to progress in this direction. This introduction was aimed at providing an outline of the theoretical constructs and issues that have emerged from this work, and their relationship to the growing corpus of experimental data—much of which is examined in detail in the remaining chapters of this volume. The most important challenge for the next phase of research will be to integrate the theoretical constructs and empirical findings that have emerged into a coherent, formally rigorous description of the mechanisms involved. The outlines of such a theory are coming into focus: Cognitive control reflects the operation of mechanisms that maintain, and appropriately update internal representations of information needed to guide processes responsible for task execution in a context-relevant, goal-directed manner. Symbolic models have provided a useful high-level description of this system. However, its implementation in the brain imbues it with capabilities (e.g., learning and inference) and constraints (e.g., a tension between generalisation and multitasking) that seem to require a finer grain of analysis and modelling. As in other domains of science, bridging these levels of analysis is a critical step towards the construction of a comprehensive theory. After half a century of research, this synthesis appears to be coming within reach, and doing so promises to provide a scientifically satisfying account of the remarkable and characteristically human capacity for cognitive control.

Acknowledgements

The author would like to thank Todd Braver, Nathaniel Daw, Ida Momennejad, Amitai Shenhav, and Tobias Egner for thoughtful comments on an earlier draft of this chapter, as well as the numerous trainees and colleagues with whom he has had the privilege to work in this rewarding area of research.

Notes

1 This idea was closely related to Broadbent's (1958) highly influential *bottleneck theory*, which asserted that attention should be thought of as a central, limited capacity filter on human information processing (the relationship between attention and control will be discussed further below).

2 This distinction parallels one between compiled versus interpreted processes in computer science.

References

Adie, W. J., & Critchley, M. (1927). Forced grasping and groping. *Brain*, *50*, 142–170.

Allport, A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology*, *24*(2), 225–235.

Allport, D. A. (1980). Attention and performance. In G. I. Claxton (Ed.), *Cognitive psychology: New directions* (pp. 112–153). London: Routledge & Kegan Paul.

Anderson, J. R.(1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *Rational analysis*. Hillsdale, NJ: Erlbaum.

Ardid, S., Wang, X.-J., & Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *Journal of Neuroscience*, *8*(32): 8486–8495.

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: Role of the subthalamic nucleus. *The Journal of Neuroscience*, *26*(9), 2424–2433.

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.

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, *73*, 640–657.

Baddeley, A. D., & Hitch, G. (1974). *The psychology of learning and memory*. New York: Academic Press.

Barlow, H. B. (1981). Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London, Series B*, *212*, 1–34.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Bianchi, L. (1895). *The functions of the frontal lobes*. Oxford: Oxford University Press.

Bickel, W. K., Jarmolowicz, D. P., Mueller, E. T., Gatchalian, K. M., & McClure, S. M. (2012). Are executive function and impulsivity antipodes? A conceptual reconstruction with special reference to addiction. *Psychopharmacology*, *221*(3), 361–387.

Bogacz, R., Brown, E. T., Moehlis, J., Hu, P., 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), 700–765.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*, 1249–1285.

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.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(2), 395–429.

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129.

Bugg, J. M., McDaniel, M. A., & Einstein, G. O. (2013). Event-based prospective remembering: An integration of prospective memory and cognitive control theories. In Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 267–282). Oxford: Oxford University Press.

Brain, W. R., & Curran, R. D. (1932). The grasp reflex of the foot. *Brain*, *55*, 347–356.

Bratman, M. (1987). Responsibility and planning. *Journal of Ethics*, *1*, 27–43.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII; Control of cognitive processes* (pp. 713–737). Cambridge, MA: MIT Press.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113.

Braver, T. S. (2015). *Motivation and cognitive control (Frontiers of Cognitive Psychology*. T. S. Braver (Ed.). Hove: Psychology Press.

Broadbent, D. E. (1958). *Perception and communication*. Elmsford, NY: Pergamon Press.
Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749.
Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 768–815.
Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*, 1462–1467.
Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930–942.
Cleermans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235–253.
Cleermans, A. (2007). Consciousness: the radical plasticity thesis. *Progress in Brain Research*, *168*, 19–33.
Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.
Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the tradeoff between exploitation and exploration. *Philosophical Transactions of the Royal Society of London Series B (Biological Sciences)*, *362*(1481), 933–942.
Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: theory and applications* (pp. 267–295). Hillsdale, NJ: Erlbaum.
Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *American Journal of Psychology*, *105*, 239–269.
Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229.
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.
Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(6), 1068–1074.
Deco, G., & Rolls, E. T. (2005). Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology*, *76*(4), 236–256.
Dehaene, S., & Changeux, J.-P. (1991). The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, *1*(1), 62–79.
Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227.
Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1–2), 1–37.
Deserno, L., Huys, Q. J., Boheme, R., Bechert, R., Heinze, H-J., Grace, A. A., … Schlagenhuaf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1595–1600.
Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.
Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems 3. *Stevens' handbook of experimental psychology: Learning, motivation and emotion* (pp. 497–534). New York: John Wiley & Sons.
Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*, 172–179.
Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*, 475–483.

Duncan, J., Seitz, R. J., Colony, J., Bor, D., Herzog, H., Ahmed, A., … Emslie, H. (2000). A neural basis for general intelligence. *Science*, *289*(5478), 457–460.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegner, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6802–6807.

Egner, T., & Hirsch, J. (2005) Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*, 1784–1790.

Einstein, G. O., & McDaniel, M. A. (2005). Prospective memory multiple retrieval processes. *Current Directions in Psychological Science*, *14*(6), 286–290.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking vs. multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective and Behavioral Neuroscience*, *14*(1), 129–146.

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, *13*, 538–539.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, *19*(8), 1120–1136.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Fuster, J. M. (1980). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe*. New York: Raven Press.

Geana, A., Wilson, R. C., Daw, N., & Cohen, J. D. (under review). Boredom, information-seeking and exploration.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Boston: MIT Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, T. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gershman, S. J., Markman, A. B., & Otto, A.R .(2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1),182–194.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*(3), 297–337.

Gittins, J. C., & Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gans (Ed.), *Progress in statistics* (pp. 241–266). Amsterdam, The Netherlands: North-Holland.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *89*(5), 881–1120.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Gollwitzer, P. M. (1996). The volitional benefits of planning. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 287–312). New York: Guilford Press.

Graziano, M., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, *2*(2), 98–113.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Gross, J. J., & Thompson, R. A. (2007). *Emotion regulation: Conceptual foundations. Handbook of emotion regulation* (pp. 3–24). New York: Guilford Press.

Haggard, P. (2008). Human volition: Towards a neuroscience of will. *Nature Reviews Neuroscience*, *9*, 934–946.

Harlow, H., & Martyn, J. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society*, *2*(3), 327–347.

Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: towards a unified developmental approach. *Developmental Science*, *13*(4), 622–635.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709.

Howard, M., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review*, *116*(4), 717–751.

Ikkai, A., & Curtis, C. E. (2011). Common neural mechanisms supporting spatial working memory, attention and motor intention. *Neuropsychologia*, *49*(6), 1428–1434.

Inzlicht, M., & Schmeichel, B. J. (2012). What Is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, *7*(5), 450–463.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale, NJ: Erlbaum.

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

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229–235.

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978) Tests of optimal sampling by foraging great tits. *Nature*, *275*, 27–31.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), 16390–16395.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology*, *8*(2), 244–259.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lhermitte, F. (1983). Utilization behavior and its relation to lesions of the frontal lobes. *Brain*, *106*, 237–255.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, *39*, 367–386.

Logan, G. D. (1994) On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 189–239). San Diego, CA: Academic Press.

Luria, A. (1966). *Higher cortical functions in man*. New York: Basic Books.

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision*, *9*(3).

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163–203.

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*, 126–135.

Marinesco, G., & Radovici, A. (1920). Sur un reflexe cutane nouveau: Reflexe palmo-mentonnier. *Revue neurologique*, *27*, 237–240.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503–507.

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 768–786.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review*, *104*(4), 749–791.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, G. A., Galanter, E., & Pribram, K. A. (1960). *Plans and the structure of behavior*. New York: Holt, Rhinehart, & Winston.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947.

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control: Three key transitions. *Current Directions in Psychological Science*, *21*, 71–77.

Musslick, S., Dey, B., Ozcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. automatic processing: A graphic-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review*, *86*, 214–255.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*, 91–94.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2013). How limited systematicity emerges: A computational cognitive neuroscience approach. In P. Calvo & J. Symons (Eds.), *The architecture of cognition: Rethinking fodor and Pylyshyn's systematicity challenge*. Cambridge: MIT Press.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience*, *27*(2), 319–333.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 358–377.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73–89.

Plate, T. A. (1995). Holographic reduced representation. *IEEE Transactions on Neural Networks*, *6*(3), 623–641.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. H. Boom & D. G. Bouwhuis (Eds.), *Attention and performance X*. Hillsdale, NJ: Lawrence Erlbaum.

Posner, M. I., & Snyder, C. R. R. (1975). In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium*. New York: Lawrence Erlbaum.

Pratt, S. C., & Sumpter, D. J. T. (2006). A tunable algorithm for collective decision-making. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(43), 15906–15910.

Ratcliff, R., & McKoon, G. (2009). The Diffusion Decision Model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*(3), 147–159.

Reynolds, J. R., Braver, T. S., Brown, J. W., & Van der Stigchel, S. (2006). Computational and neural mechanisms of task switching. *Neurocomputing*, *69*(10–12), 1332–1336.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*, 585–590.

Roca, M., Torralva, T., Gleichgerrcht, E., Woolgar, A., Thompson, R., Duncan, J., & Manes, F. (2011). The role of area 10 (BA10) in human multitasking and in social cognition: A lesion study. *Neuropsychologia*, *49*, 3525–3531.

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

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

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*(1), 101–130.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychological Review*, *84*(2), 1–66.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science*, *12*, 101–108.

Schurger, A., Pereira, F., Treisman, A., & Cohen, J. D. (2010). Reproducibility of activity characterizes conscious from non-conscious neural representations. *Science*, *327*(5961), 97–99.

Shaffer, L. H. (1975). Multiple attention in continuous verbal tasks. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance* (pp. 157–167). New York: Academic Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B*, *298*, 199–209.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 217–240.

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.

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, *5*, 85–90.

Simen, P., Contreras-Ros, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: Empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1865–1897.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

Simon, H. A. (1992). What is an 'explanation' of behavior? *Psychological Science*, *3*, 150–161.

Simoncelli, E. P., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Susillo, D. (2014). Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology*, *25*, 156–163.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proceedings of the Seventh International Conference on Machine Learning*.

Tanner, W. P. J., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, K. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309–318.

Todd, M., Niv, Y., & Cohen, J. D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. *Advances in Neural Information Processing Systems (Vol. 20)*. Cambridge, MA: MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tombu, M. N., Asplund, C. L., Dux, P. E., Godwin, D., Martin, J. W., & Marois, R. (2011). A unified attentional bottleneck in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 13426–13431.

Townsend, J. T. (1972). A note on the identifiability of parallel and serial processes *Perception & Psychophysics*, *10*(3), 161–163.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1–25.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A.,& Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447.

Usher, M., Cohen, J. D., Haarmann, H., & Horn, D. (2001). Neural mechanism for the magical number 4: Competitive interactions and nonlinear oscillation. *Behavioral and Brain Sciences*, *24*(1), 151–152.

Vargas, M. (2013). *Building better beings*. Oxford: Oxford University Press.

Verbruggen, F., & Logan, G. D. (2008). Automatic and controlled response inhibition: Associative learning in the go/no-go and stop-signal paradigms. *Journal of Experimental Psychology: General*, *137*(4), 649–672.

Venkatraman, V., Rosati, A. G., Taren, A. A., & Huettel, S. A. (2009). Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. *Journal of Neuroscience*, *29*, 13158–13164.

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, *115*(2), 518–525.

Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: a neurocomputational model. *Frontiers in Behavioral Neuroscience*, *9*.

Watkinson, S. C., Boddy, L., Burton, K., Darrah, P. R., Eastwood, D., Fricker, M. D., & Tlalka, M. (2005). New approaches to investigating the function of mycelial networks. *Mycologist*, *19*, 11–17.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2): 395–415.

Wickens, D. D. (1984). Processing resources in attention. In R. Parasuraman, D. R. Davies, & J. Beatty (Eds.), *Varieties of attention* (pp. 63–102). New York: Academic Press.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*(2), 329–355.

Wiener, N. (1948). *Cybernetics, or control and communication in the animal and the machine*. Cambridge: MIT Press.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (in press). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Attentional control: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 700–717.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, *3*(2), 179–193.