Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors

S. F. Feng · M. Schwemmer · S. J. Gershman · J. D. Cohen

Abstract Why is it that behaviors that rely on control, so striking in their diversity and flexibility, are also subject to such striking limitations? Typically, people cannot engage in more than a few—and usually only a single—control-demanding task at a time. This limitation was a defining element in the earliest conceptualizations of controlled processing; it

Relevance of this article to the commemoration of Ed Smith The span of Ed Smith's work, as well as its influence on cognitive science and cognitive neuroscience, has been both deep and broad, embracing both experimental work and formal theory. Some of his earliest and most influential work focused on the nature of similarity relationships and conceptual representations. Later, in his collaboration with John Jonides, he shifted his interest to cognitive control and, in particular, how it is implemented in the brain. It seems fitting, therefore, that the work that we present in this special issue bridges these two themes, relating a fundamental functional characteristic of controlled processing—its limited ability for multitasking—to the nature of representations and processing in the brain. Although our contribution focuses on capacity constraints in cognition, the breadth and force of Ed's work are evidence of the truly awe-inspiring achievements possible even within the bounds of those limitations.

S. F. Feng · M. Schwemmer · J. D. Cohen
Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey, USA

S. F. Feng · M. Schwemmer
Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey, USA

S. J. Gershman
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

J. D. Cohen (✉)
Department of Psychology, Princeton Neuroscience Institute, Princeton, NJ 08544, USA
e-mail: jdc@princeton.edu

remains one of the most widely accepted axioms of cognitive psychology, and is even the basis for some laws (e.g., against the use of mobile devices while driving). Remarkably, however, the source of this limitation is still not understood. Here, we examine one potential source of this limitation, in terms of a trade-off between the flexibility and efficiency of representation ("multiplexing") and the simultaneous engagement of different processing pathways ("multitasking"). We show that even a modest amount of multiplexing rapidly introduces cross-talk among processing pathways, thereby constraining the number that can be productively engaged at once. We propose that, given the large number of advantages of efficient coding, the human brain has favored this over the capacity for multitasking of control-demanding processes.

Keywords Computational model · Capacity constraint · Decision-making

Introduction

The human ability for controlled processing is striking in its power and flexibility, and yet equally striking in its limitations. People are famously poor at multitasking control-demanding behaviors; they are often able to execute only a few, and sometimes no more than one at a time, and these limitations are central to debates that have broad social significance (e.g., the use cell phones while driving). Nevertheless, there is still little agreement about the source of these limitations.

The limits in our ability to multitask control-demanding behaviors have sometimes been taken to suggest that the capacity for control itself is limited. Indeed, limited capacity was a defining feature in the earliest conceptualizations of controlled processing (e.g., Posner & Snyder, 1975; Shiffrin & Schneider, 1977), which asserted that controlled processing

relies on a single, centralized, and capacity-limited general purpose resource. Although early theories failed to specify the nature of this resource, subsequent theories have been more explicit about the mechanisms involved (e.g., Anderson, 1983; Baddeley, 1986; Engle, 2002; Just & Carpenter, 1992)—for example, linking the constraints on control to the long recognized capacity limits of working memory (e.g., G.A. Miller, 1956; Oberauer & Kliegl, 2006; Sternberg, 1966) and/or to the constraints on centralized procedural resources needed to execute task-related processes (e.g., Salvucci & Taatgen, 2008). However, these ideas risk stipulating rather than explaining the limitation. Furthermore, the notion of a resource limitation is startling, if one considers that control processes are known to rely on the function of the prefrontal cortex, a structure that occupies a major fraction of the human neocortex and comprises about three billion neurons (Herculano-Houzel 2009; Pakkenberg & Gundersen, 1997; Thune, Uylings, & Pakkenberg, 2001; Williams & Herrup, 1988). Thus, it seems unlikely that a structural limitation alone can explain the constraints on multitasking. Though some have proposed that the constraints reflect metabolic limitations (Muraven, Tice, & Baumeister, 1998), this too seems unlikely, since other functions, such as vision, routinely engage widespread regions of neocortex in an intense and sustained manner.

Multitasking versus multiplexing

An alternative approach to understanding limitations in control-dependent multitasking is to consider that these may arise from functional (i.e., computational) rather than, or perhaps in addition to, structural constraints. One important functional constraint may be the problem presented by multiplexing of representations—that is, the use of the same representations for different purposes by multiple processes.[1] This idea has its roots in early "multiple-resource" theories of attention (Navon & Gopher, 1979; see also Allport, 1980; Logan, 1985; Wickens, 1984). These argued that the degradation in behavior observed when people try to multitask may be due to cross-talk within local processing resources specific to the particular tasks involved, rather than to the constrained resource of a single, centralized general-purpose control mechanism (Baddeley, 1970; Posner & Snyder, 1975). Cross-talk arises when two (or more) tasks make simultaneous demands on the same processing or representational apparatus. Insofar as each is capable of supporting only a single process or representation at a given time, these can be thought of as limited-capacity "resources," though multiple different

resources are assumed to support the various types of processes and representations of which the system is capable. Accordingly, cross-talk can arise anywhere in the system where two tasks make competing use of the same local resource(s). This, in turn, will interfere with multitasking.[2]

A classic example contrasts two dual-task conditions: echoing an auditory stream while simultaneously typing visually presented text (copy-typing), versus simultaneously reading aloud and taking dictation. The former pair is relatively easy to learn, whereas the latter is considerably more difficult (Shaffer, 1975). The multiple-resource explanation suggests that echoing and copy-typing involve nonoverlapping representations and processing pathways (one auditory–phonological–verbal, the other visual–orthographic–manual). Because they make use of distinct resources, there is no risk of cross-talk, so it is possible to do both at once (see Fig. 1A). In contrast, reading out loud and dictation make shared use of both phonological and orthographic representations (see Fig. 1B), and thus are subject to the problem of cross-talk: The two tasks demand that different representations of a given type be activated at the same time (e.g., the one to be read and the other to be transcribed). Insofar as this is not possible (e.g., to simultaneously activate phonological representations corresponding to different stimuli), competition will arise that, if not properly managed, will lead to interference and degradation of performance. Allport , Antonis, and Reynolds (1972) summarized this clearly in one of the earliest articulations of the multiple-resource (or "channels") hypothesis:

> In general, we suggest, any complex task will depend on the operation of a number of independent, specialized processors, many of which may be common to other tasks. To the extent to which the same processors are involved in any two particular tasks, truly simultaneous performance of these two tasks will be impossible. On the other hand, the same tasks paired, respectively with another task requiring none of the same basic processors can in principle be performed in parallel with the latter without mutual interference. (p. 233)

The problem of mutual interference, or cross-talk, arises due to the multiplexing of representations—that is, to the use of the same representations for different purposes—for

---

[1] The term "multiplexing" is often used in the context of signal processing to refer to the simultaneous communication of two or more signals over the same processing channel. Here, we use it in a broader sense, to refer to the allocation of a single mechanism (or set of representations) to a multiplicity of purposes.

[2] Note that our focus here is on *concurrent* multitasking, sometimes referred to as "perfect timesharing" or "pure parallel processing." This contrasts with forms of multitasking in which the processes associated with the different tasks are interleaved, with only one process being executed at a time under the control of a central processor and its associated scheduler. Such multitasking may approximate concurrency at a coarser level (as, for example, often occurs in computers). However, the limits to performance are typically attributable to the centralized processor and/or its scheduler, and not to direct interaction between the tasks themselves (though, in some systems, bottlenecks can arise for either reason; see, e.g., Salvucci & Taatgen, 2008).
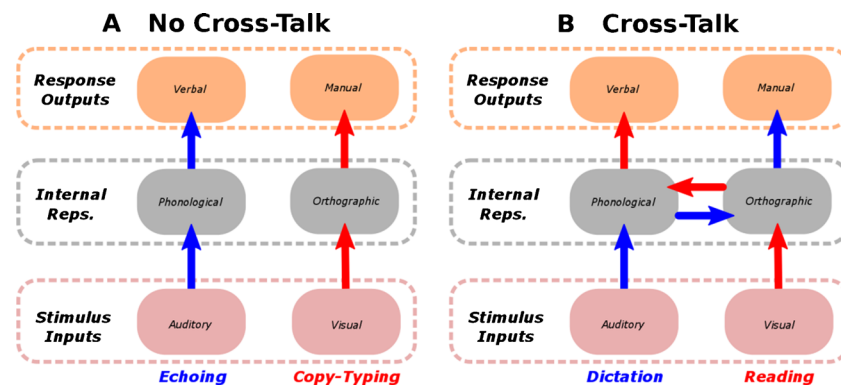
Fig. 1 Multitasking and cross-talk: Schematic diagrams of the processing pathways involved in two dual-task situations. Panel **A** shows the pathways engaged by verbally repeating auditorily presented words (echoing) and transcribing visually presented text (copy-typing). Since the pathways carrying the information relevant to each task from input to output do not overlap, there is no cross-talk, making it possible to carry out both tasks simultaneously (i.e., to multitask). Panel **B** shows the pathways engaged by typing auditorily presented words (dictation, blue arrows in the online version) and reading text out loud (red arrows in the online version). In this case, the pathways for the two tasks overlap; that is, they make common use of the same sets of representations (e.g., phonological representations are needed for both encoding the words for dictation and generating the responses for reading). Insofar as each task may involve representing different specific information items of the same type, the pathway overlap will produce cross-talk and interference, making it difficult, if not impossible, to multitask

example, the use of phonological representations for both encoding auditorily presented words and reading words out loud. Presumably, multiplexing occurs in the brain because it is efficient and affords flexibility. For example, we would not want to have two sets of phonological representations, one for listening and another for speaking, as well as additional ones for echoing, singing, and so forth. Furthermore, multiplexing allows representations that arise for one purpose to be put to use for others, and can also support powerful forms of computation, such as similarity-based inference (Forbus, Gentner, & Law, 1995; Hinton, McClelland, & Rumelhart, 1986). Theories of "embodied" or "grounded" cognition (e.g., Barsalou, 2008) are based on similar arguments, proposing that the same neural apparatus used for sensory–motor processes is also used for more abstract forms of thought. In support of this, neuroimaging studies have documented the use of the specific parts of the visual system for both perception and imagery (Kosslyn, Thompson, & Alpert, 1997), and the motor system for action execution, planning, and even interpretation (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese & Goldman, 1998; Goebel, Khorram-Sefat, Muckli, Hacker, & Singer, 1998; Jeannerod, 1994).

As we noted above, however, multiplexing carries with it the problem of cross-talk. Furthermore, this problem is not limited to competition between instructed processes. It extends more generally to situations in which interference can arise from competing habitual (or "automatic") processes. In general, it has been argued that a primary purpose of control is the management of such cross-talk (e.g., Berlyne, 1957; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Meyer & Kieras, 1997), by restricting the number of competing processes that are executed at once. Note, however, that according to this account, limitations arise from cross-talk within

local, task-specific "resources" (e.g., phonological or visual representations), rather than from the capacity constraint on a single, centralized control resource (such as working memory). Indeed, from this perspective, control helps *solve* the problem of cross-talk. However, its consistent engagement under circumstances that involve cross-talk (and, therefore, its close association with them) could engender the misinterpretation that control itself is fundamentally subject to (rather than a solution for) capacity constraints on performance.

Debate is still ongoing about whether limits in the ability to multitask are due to the capacity constraints of a centralized control mechanism (perhaps even for the very same reasons that they arise in more specialized mechanisms), or whether they can be accounted for entirely in terms of the cross-talk among task-specific mechanisms, and this debate has fueled considerable computational modeling work (e.g., Meyer & Kieras, 1997; Salvucci & Taatgen, 2008). However, both types of account would seem to concur on two fundamental points regarding the importance of cross-talk: (1) It represents a critical factor in limiting multitasking (wherever it may occur), and (2) the need to limit it drives the need for control in the first place. Thus, achieving a deeper understanding of the relationship between cross-talk, control, and multitasking is an important goal, especially if general properties can be identified. The work that we present here is an attempt to address this goal.

We conducted a quantitative examination of the influence that multiplexing (and associated cross-talk) has on control and multitasking. In striving for generality, we did this under a simple and generic set of conditions: Each task involved a direct set of mappings from input to output; cross-talk occurred when tasks involved the same inputs or outputs; control acted only to select tasks for execution; and no *intrinsic* capacity constraints were placed on control. The last of these conditions was meant to isolate the influence that multiplexing

had on multitasking, independent of any additional influence that constraints on the capacity for control itself might have. We used two different formalisms to implement these conditions, and in each instance optimized the processing parameters (including control) to maximize aggregate performance over all tasks. We then examined the extent to which the introduction of multiplexing impacted the number of tasks that could be executed at once. Not surprisingly, we found that the optimal number of tasks to carry out decreased as multiplexing increased. Critically, however, this decrease was precipitous and extreme, and was virtually unaffected by the size of the system. This suggests the intriguing idea that the presence of multiplexing—and the potential for cross-talk that it introduces—not only poses a demand for control, but at the same time imposes limits on the amount of control that should be allocated. This, in turn, suggests that the extensive use of multiplexed representations in the brain may provide at least one normative source of constraint on the multitasking of control-dependent processes.

In the sections that follow, we provide a brief review of previous modeling work addressing the relationship between cross-talk and control, which formed the basis for the present work. We then describe extensions of this work, using two models to quantify the relationship between multiplexing, cross-talk, and limits in the multitasking of control-dependent processes. Finally, we conclude with a discussion of how our findings relate to other accounts of constraints on the capacity for control-dependent processing, as well as of promising directions for future work on this topic.

A simple model of controlled processing and cross-talk

We start with a model of the Stroop task (MacLeod, 1991; Stroop, 1935). Although this task is not typically thought of in the context of multitasking, it provides perhaps the simplest example of the problem that multiplexing and cross-talk present to multitasking, and it is one of the most intensively studied examples of controlled processing. In this task, participants are presented with a written word displayed in color (e.g., GREEN presented in red) and are asked to name the color in which it is displayed (e.g., to say "red"). This invariably takes longer (and is more prone to error) than if they are simply asked to read the word, or to name the color of a word that refers to the same color (e.g., RED presented in red). The traditional interpretation of these findings—which are among the most robust in all of behavioral science—is that color naming is a control-demanding process (slower, effortful, and subject to interference), whereas word reading is automatic (faster, effortless, and performed even in the absence of control; Posner & Snyder, 1975). A more recent and nuanced interpretation is that the demands for control are context dependent, and in particular, dependent on the nature of any

competing tasks (e.g., Cohen, Dunbar, & McClelland, 1990; Engle, 2002; Kahneman & Treisman, 1984). By all accounts, however, when a task involves behavior that must compete with other—more reflexive, habitual, or otherwise prepotent—responses, then performance on the task will demand the allocation of control in order to avert the cross-talk induced by the more automatic process. Color naming in the Stroop task provides a clear of example of this.

Cohen et al. (1990) described a neural-network model that provided a mechanistic account for many of the findings observed in the Stroop task. In that model (Fig. 2A), task performance was simulated as the flow of activity between sets of processing units that comprised pathways from stimuli to their responses. The model was made up of two sets of input units: one representing the orthographic form of the stimulus, and the other its color. Each of these projected to a corresponding set of associative ("hidden") units that, in turn, converged on a set of response units representing the verbal (phonological form of the) response. Thus, Cohen et al. (1990) described two pathways: one for word reading, composed of connections from the orthographic input units to their associative units and from them to the verbal response layer, and one for color naming (from the color inputs units, by way of their associative units, to the response layer). Note that the response units served two purposes—reading the word and naming the color of the stimulus. In this respect, the representations in the response layer were multiplexed, and thus constituted a limited "resource" for which the two pathways competed.

To simulate the greater automaticity of word reading, the connections along that pathway were stronger than those along the color-naming pathway. As a consequence, when a conflict stimulus was presented (e.g., GREEN presented in red), the model responded to the word rather than the color, much as participants do when they are presented with such stimuli and provided only with the instruction to "respond" (i.e., without specifying which task to perform). However, when they are instructed to name the color, people can do so. This is a clear example of the capacity for control—that is, the ability to respond to the weaker dimension of a stimulus, in the face of competition from a stronger, prepotent response. The model explained this in terms of the "top-down" influence of a set of control units used to represent the task to be performed. Activating the color-naming task unit sent additional biasing activity to the associative units in the color pathway, so that they were more responsive to their inputs (see Cohen et al., 1990 or Cohen et al., 2004 for fuller explanations of this effect). In this way, the model was able to respond to the color of the stimulus, even in the face of competition from the otherwise stronger word-reading pathway. This illustrated a fundamental function of control: the management of cross-talk in the service of task performance.

This model and similar ones have been used to account for performance in a wide range of cognitive tasks that rely on

Fig. 2 Stroop model. (**A**) Original model of the Stroop task (Cohen et al., ▶ 1990). The bottom units are used to represent stimuli along the two dimensions (color and orthographic form). These pass activity to intermediate (associative) units (middle layer) that, in turn, pass activity to output units (top layer) that represent the possible verbal responses. The activities of the intermediate units are modulated by the task demand (control) units (upper right box) that represent each of the two possible tasks (color naming and word reading). Some links between units (brown in the online version) represent excitatory connections (positive weights); other links represent inhibitory connections (negative weights; red in online version); and thicker lines represent stronger connections. The word-reading pathway has stronger connections than does the color-naming pathway (presumably due to greater practice and/or more consistent mapping), which implements the greater automaticity of this task. As a result, in order to perform the color-naming task and respond correctly to an incongruent stimulus (e.g., GREEN presented in red), the color-naming control unit must be activated (see the text for explanation). In this respect, color naming is a control-demanding task, relying on the corresponding control unit to manage potential interference from cross-talk produced by the word-reading pathway. (**B**) Extended model that includes a pathway for the word–action task described in the text. Note that multiplexing now takes place in both the verbal response layer and the intermediate, word-processing layer. As a consequence of this, color-naming and word–action tasks cannot be performed simultaneously without the risk of cross-talk

A Stroop Model

B Extended Model

control (e.g., Botvinick & Cohen, in press; Braver & Cohen, 2000; Cohen et al., 1990; Liu, Holmes, & Cohen, 2008; O'Reilly, Noelle, Braver, & Cohen, 2002; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; Servan-Schreiber, Bruno, Carter, & Cohen, 1998; Yu, Dayan & Cohen 2009) and for the role of neural structures (such as the prefrontal cortex, basal ganglia, and dopamine systems) in supporting this function (Braver & Cohen, 2000; E. K. Miller & Cohen, 2001; O'Reilly & Frank, 2006). More specifically, such models have also directly addressed the role that control plays in managing cross-talk (e.g., Botvinick et al. 2001; Shenhav, Botvinick, & Cohen, 2013; Yeung, Botvinick, & Cohen, 2004). However, their focus has been on situations in which control is needed to manage cross-talk from unintended processes, and not on situations in which cross-talk arises from multitasking (i.e., from other *intended* processes). Accordingly, the models have not directly addressed limitations in the multitasking of control-dependent processes. However, an important source of such limitations is implicit in these models.

Cross-talk and limits on multitasking

Imagine an extension of the Stroop task in which a person is asked not only to name the color of the word, but also to respond to the word itself with some other action (e.g., pressing a different button for each word). Figure 2B shows an elaboration of the Stroop model corresponding to this set of tasks, in which the word units are now mapped to a set of manual actions, in addition to verbal responses. Carrying out both tasks simultaneously requires activating both the color and word-reading units. Although in principle this is
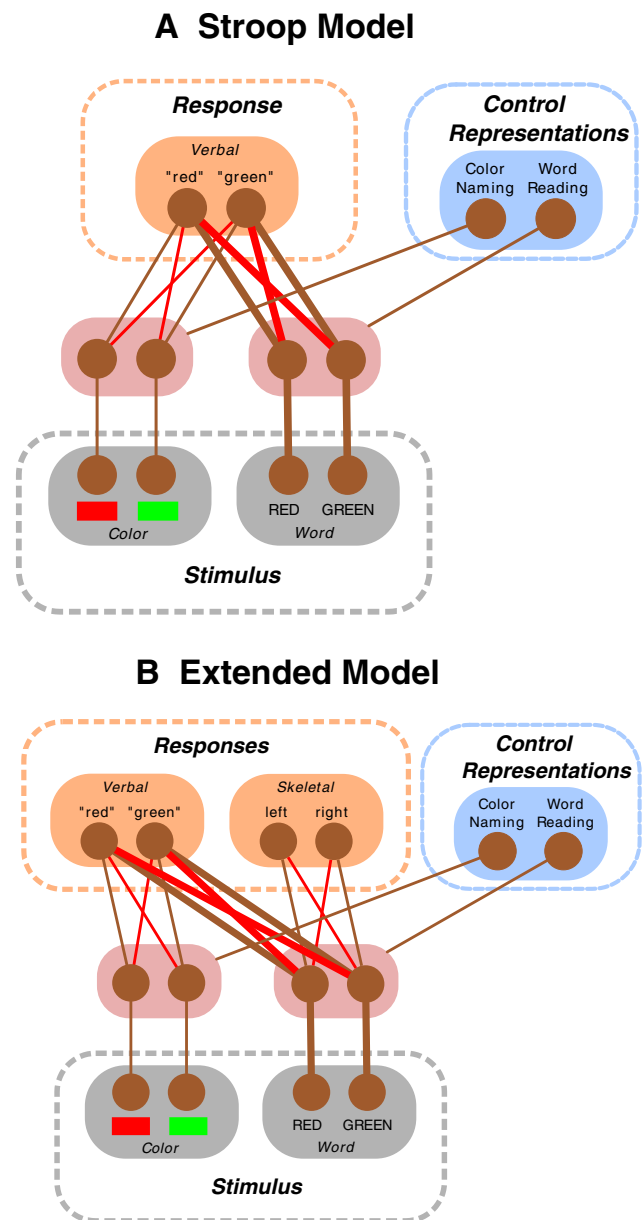
possible—there are no intrinsic constraints in the model on how many control units can be activated—the consequence is that color naming will once again be subject to cross-talk from the word. Attempting to perform these two tasks makes it clear that it is almost (if not entirely) impossible to do them at the same time, much like reading and dictation.[3] The original Stroop model shows how multiplexing presents a risk of

---

[3] Note that it may be possible to learn to do these tasks simultaneously. However, this would presumably require the allocation of a new set of intermediate units, dedicated to the mapping from words to actions, as well as the associated control representations, and would come at the cost of considerable practice, as well as reduced coding efficiency and the loss of any other benefits that accrue from multiplexing (e.g., inference).

cross-talk, and thus poses the need for control. This extension shows that multiplexing can also impose a limit on the amount of control that should be allocated at once.

In the original Stroop model, cross-talk occurred due to multiplexing at the response layer, where the same verbal responses are afforded to two potentially competing processes—color naming and word reading. More generally, however, such cross-talk can arise anywhere that there are multiplexed representations (such as the associative units in the extended Stroop model, or the examples of reading and dictation described above). To the extent that multiplexing is widespread, the potential for cross-talk and the corresponding need for control will be high. At the same time, this may bring with it constraints on the amount of control that should be allocated. As an analogy, consider the problem faced in designing a network of train tracks and then scheduling the trains: Increasing the directness of connections between sources and destinations necessarily increases the density of tracks, and therefore the number of crossing points. Managing these, in turn, requires limiting the number of trains that are running at the same time, or carefully scheduling their transit (e.g., by restricting the number of green lights that can be simultaneously lit at crossing points). By analogy, dense overlap of representations in the brain (multiplexing) may be an efficient way of encoding information (and serving other functions, such as constraint satisfaction and inference); however, this brings with it the potential for interference due to cross-talk, and thus may restrict the number of control-dependent processes that can safely be licensed at once, and thereby limit multitasking.

The foregoing provides a qualitative account of why the multiplexing of representations introduces the need for control, and the concomitant constraints that it places on multitasking. However, it does not provide a quantitative account. For example, how severely does the presence of multiplexing constrain multitasking, and to what extent are these effects scale-invariant—that is, to what extent do they depend on the size of the system? To address these questions, we constructed variants of the Stroop model that allowed us to amplify the number of pathways, manipulate the amount of multiplexing, and normatively evaluate the number of pathways that could simultaneously be engaged—that is, the limits of multitasking. We hypothesized that introducing even modest amounts of multiplexing into a network would dramatically constrain its ability to multitask, and that this effect would be largely insensitive to the size of the network; that is, improvements in multitasking ability would be limited, even with large increases in network size. The latter case is important, given the striking disparity between the large number of potential processing pathways in the human brain, yet the severely restricted number of control-demanding tasks that people can simultaneously perform.

Methods

We constructed two variants of the Stroop model that extended it to simulate more than two pathways. These used the same basic architecture, but different processing functions and objective functions (in order to evaluate the influences of cross-talk and control on performance). The first model (I/O matching) used a simple linear mapping from inputs to outputs for each pathway, and the match between the output and a specified target value as the objective function. The second model (DD) used a drift diffusion process to simulate the dynamics of processing in each pathway, and the reward rate as the objective function. We present both models, in order to establish the generality of our findings and extrapolate the anticipated influence of critical factors (such as linearity of the processing function) to a broader class of models.

In the section that follows, we describe the overall network architecture that was common to both models, followed by the processing characteristics and objective functions that were specific to each. We then describe the procedures used to optimize processing and control in each network, and to evaluate the corresponding limits of multitasking. Finally, we describe the three factors that we manipulated to determine the influence of multiplexing on the limits of multitasking.

Network architecture

Pathways

Figure 3 shows the overall network architecture used for both models. For simplicity (and analytic tractability, in the case of the DD model), we restricted these pathways to simulating simple two-alternative forced choice tasks. All of the units in the models used linear processing functions (relating their input to their output), and the models were composed of only input and output layers. We refer to the projections from each pair of input units to a pair of output units as a pathway P, with each pathway implementing the input–output mapping corresponding to a different task. Within each pathway, each input unit of the pair was connected to one of the output units in the corresponding pair with a weight of .5, and the other output unit with a weight of –.5 (these specific values, and those for the activity of the input units described below, were used to simplify the mathematical analysis—see SOM A for details). Thus, all weights were symmetric within a pathway, and their magnitudes were equal across all pathways. The weights of connections between input units and output units that did not form a pathway were set to 0.

For the purposes of analysis, we focused on a subset of pathways relevant to the tasks to be performed (P_{rel})—that is, the tasks desired to be multitasked. We denote the number of relevant pathways by N. In Fig. 3, these are illustrated as the
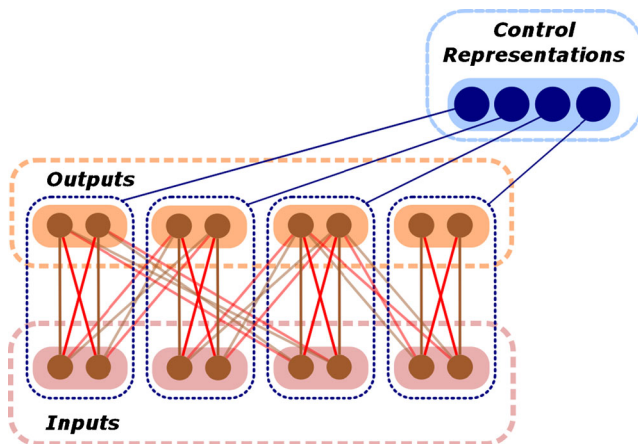
Fig. 3 Network architecture. The example shows a network with four pairs of input and output units ($N=4$), with overlapping pathways ($F=1$) and the level of incongruence used in most simulations ($\varphi=.75$). Pairs of input units connected to directly overlying pairs of output units (i.e., those contained within the same dark-dotted box) constitute relevant pathways—that is, those to be multitasked. Connections between pairs of input and output units in different boxes (shown in lighter colors) represent irrelevant pathways. Some connections have positive, and others negative, weights, colored as in Fig. 2. Connections between pairs of input and output units that have weights of the same sign are congruent, and those having opposite signs are incongruent (see note 6 for an explanation). Control units are shown in the upper right box. Their connections to the corresponding pathways are also shown in darker color, indicating that they have a multiplicative effect on the activity of the input and output units in those pathways (see the text for an explanation)

pathways that connect a pair of input units to the immediately overlying pair of output units. The remaining pathways (P_{irrel}, shown as crossed pathways in Fig. 3) implemented multiplexing in the model—that is, the assumption that other, overlapping input–output mappings subserve tasks that are useful in other contexts but are irrelevant in the present one—and therefore introduced the potential for cross-talk. Two factors determined the extent of this cross-talk: the prevalence of pathway overlap (multiplexing) and whether inputs to overlapping pathways were likely to generate the same or different (and therefore interfering) responses. We discuss these factors below, in the Network Configuration section.

Control

Each pathway in the network was subject to control. As in the Stroop model, this was implemented by a dedicated unit that modulated the flow of activity from the input to the output units. Thus, N control units were assumed, the activities of which were defined by a vector $\kappa = (k_1, k_2, \ldots, k_N)$. In the Stroop model, control signals were implemented as a bias added to the input of the associative units in the corresponding pathway. Because units in that model used nonlinear (sigmoidal) processing functions, the bias had the effect of modulating the gain of those units (Cohen et al., 1990).

However, units in the present model used linear processing functions, so we implemented modulation as a direct multiplicative effect. Specifically, the control signal k_i for a given pathway multiplied the activity of the units in that pathway. The values of k were bounded by 1 ($k_i \in [0, 1]$), paralleling the bounded effects of control in the case of nonlinear units (i.e., bounds on the derivative of the sigmoidal function). Furthermore, control was applied to both the input and output units in each relevant pathway, in accord with the idea that these implemented the input–output mapping corresponding to the task to be controlled.[4] Finally, it should be noted that no intrinsic constraints were placed on the number of control units that could be activated at a given time. This was determined exclusively by procedures that optimized the overall performance of the network, as we describe below.

Processing

I/O-matching model This model implemented a simple, single-pass, feed-forward network composed of linear output units. On each trial, one unit for each pair of input units in a pathway was assigned an activity value i of .5, and the other a value of –.5 (these can be thought of as corresponding to the net input to the intermediate units in the Stroop model in Fig. 2A; Cohen et al., 1990). For simplicity in evaluating the effects of cross-talk (discussed under Network Configuration below), and without loss of generality, the leftmost input unit was always assigned a value of .5 and the rightmost unit –.5. The activity of each output unit was then computed as the sum of all of the inputs connected to it, scaled by both the connection weight and each pathway's control value. For example, the activity of the qth relevant pathway's left output unit (indexed as "1," where the right unit was indexed as "2") was

$$Y_{q1} = \sum_{p \in P_q} k_p \left(w_{p1} i_{p1} + w_{p2} i_{p2} \right), \qquad (1)$$

where the sum is over all connected pathways to the qth pair of output units, $[i_{p1}, i_{p2}]$ is the activity of the left and right input units of the pathway p, $[w_{p1}, w_{p2}]$ are the weights of the corresponding connections (as described above), and k_p is the value of control for the pathway that includes these units. For the purposes of the objective function (see below), we constrained the activity of all output units to remain in the range $[-1, +1]$. We did this by further multiplying all weights w_j by $1/R$, where R was the total number of projections

---

[4] The same assumption was implicit in the Stroop model. Although in that model the control (task demand) units projected only to the associative units, the response units were also assumed to be subject to some form of modulation (e.g., see Simulation 6 and Fig. 2B in Cohen et al., 1990).

received by each output unit ($R = F + 1$, where F was the parameter that determined the degree of pathway overlap—see the Pathway Overlap section below).

The "task" to be performed by each pathway in the P_{rel} set was simply for the output units in that pathway to match the activity of the corresponding input units. Accordingly, the performance criterion (i.e., objective function) used for this model was the sum of the squared differences between each input–output pair, divided by the number of pathways in the P_{rel} set—that is, the mean squared error (MSE) of the output vector as compared to the input vector. However, in order to account for control on the output units, a subtle but important consideration also had to be taken into account: the influence of pathways assigned low control values. As we will discuss below, optimization procedures reduced control to pathways that were subject to cross-talk, in order to diminish their deleterious contribution to network performance. However, if left unchecked, the outputs of a fully inactivated pathway might still be heavily influenced by cross-talk from irrelevant pathways. To properly account for this, we scaled the values of the output units (as well as the input units) by the corresponding control value, so that the scaled activity modifying Eq. 1 above was $\overline{Y}_{q2} = k_q Y_{q2}$. These scaled outputs are what were matched to the inputs in the objective function (see SOM A for details). In this way, control influences both the outputs and inputs in a relevant pathway. Note that as constructed, the objective function favors allocating control to (i.e., activating) as many pathways as possible (i.e., maximizing multitasking), while weighing this against the costs to performance of any interference introduced by cross-talk.

DD model This model used the same network architecture as the I/O-matching model. However, the activity of the output units was determined by simulating the dynamics of information accumulation using a drift diffusion process (Ratcliff, 1978; Ratcliff, Van Zandt, & McKoon, 1999) for each pathway. Drift diffusion processes have been used widely to simulate performance in two alternative forced choice decision tasks (Balci et al., 2011; Ratcliff & Rouder, 1998; Simen et al., 2009), and have been shown to provide a good approximation of underlying neural processes (Gold & Shadlen, 2007; Schall, 2001; Shadlen & Newsome, 2001; Usher & McClelland, 2001). Furthermore, they have the advantage of being tractable to analysis, and thus can be parameterized for optimal performance (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; we will return to this below). Here, we extended the standard implementation of a DD model to include multiple overlapping processes. Thus, the activity of each output unit was a linear combination of its inputs, as in Eq. 1. However, in this case, the difference in activity between the

two output units in a pathway then served as the drift rate for a DD process, and a response was made when this crossed a specified threshold.[5] The drift rate was modulated not only by activity of the input units and the value of the control parameter k_p for that pathway (i.e., for pathways in the P_{rel} set), but also by the input received from, and control parameters for, overlapping pathways (i.e., in the P_{irrel} set; see Fig. 3 and SOM A for details).

To insure that performance was sampled over as representative a range of input conditions as possible, the activity of each input unit on each trial was drawn from a uniform distribution of values between 0 and 1. This contrasted with the I/O-matching model, in which inputs were either .5 or –.5. For the DD model, it was important to sample inputs of varying strengths since the performance criterion used for this model (reward rate—see below) depended non-linearly on drift rate (and therefore input activity; see Bogacz et al., 2006). For completeness, in addition to the uniform distribution, we also examined other distributions of input values and obtained similar results (see SOM B). As in the I/O-matching model, greater activity was always assigned to the left versus right input unit in each pair.

In accord with prior optimality analyses of DD processes (e.g., Bogacz et al., 2006), we used reward rate as the objective function for this model (see SOM A). Specifically, we optimized the response threshold for the DD process implemented by each pathway for the conditions of each trial (see the Optimization section, below), so as to maximize its reward rate on that trial. Then, paralleling the I/O-matching model, we scaled the reward rate for each pathway by its corresponding level of control. Finally we summed the scaled reward rates for each pathway to obtain a reward rate for the entire network:

$$RR = \sum_{p=1}^{N} k_p RR_p, \qquad (2)$$

Here, RR_p denotes the reward rate for each pathway. Thus, each term in the sum of Eq. 2 represents the pth pathway's *reward rate,* scaled by its control policy κ, and the sum is taken over all pathways. As with the I/O-matching model, the objective function favored allocating control to as many pathways as possible (maximizing multitasking), subject to the constraint that doing so did not degrade performance (and thereby decrease reward rate).

Optimization of performance and control policy

For each of the two models described above, we sought to identify the optimal allocation of control over the pathways in

---

[5] This is very similar to the implementation of the response mechanism in the Stroop model (Cohen et al., 1990). Although that model used nonlinear processing units, the response was determined by a linear combination of the activity of the output units, which implemented a drift diffusion process.

the P_{rel} set (i.e., how much multitasking could be supported). We examined how this was influenced by pathway overlap (that is, the degree of multiplexing in the network). As we described above, the objective function for both models favored allocation of control to as many pathways as possible (while remaining sensitive to any degradation in performance introduced by cross-talk). This insured that our analyses favored multitasking as much as possible, and therefore that estimates of the limitations in this ability were as conservative as possible.

In order to further insure that optimization of control policy favored multitasking as much as possible, we needed to be sure that processing in each pathway in the P_{rel} set was optimized for each trial, given the inputs and allocation of control on that trial. This was not a problem for the I/O-matching model, since the outputs were a simple linear sum of the inputs modulated by control. However, insuring optimal performance for the DD model required an additional step, since its performance depended not only on the inputs and control for each pathway, but also on the response thresholds. As described in Bogacz et al. (2006), in the DD model a single response threshold optimizes performance for a given drift rate (determined, in our model, by the inputs and control in a pathway). Thus, to insure that performance was optimized for the current control policy, it was necessary to optimize the threshold in each pathway for the inputs and amount of control allocated to it. We did this using the analyses reported in Bogacz et al. (2006).

Specifically, we used the following iterative procedure to identify the optimal control policy for a given network architecture. First, we sampled a set of inputs and arbitrarily chose a level of control for each pathway in the P_{rel} set. For the I/O-matching model, we then computed the outputs and evaluated the objective function (see Eq. S1 in SOM A). For the DD model, we identified the optimal response threshold for each pathway on the basis of its inputs and allocation of control, and then analytically computed the resulting reward rates for each output unit's DD process using equations in Bogacz et al. (2006). For both models, we then used parallelized optimization software to iteratively sample control policies, and identify the one that corresponded to a global maximum of the corresponding objective functions (see SOM C for details). In the results presented below, each point was obtained by solving the relevant optimization problem 10,000 times.

Evaluation of limits on multitasking

The purpose of the procedures described above were to determine, for a given network configuration, the optimal control policy for that configuration. We then identified the degree of

multitasking that this afforded. That is, our primary goal was to determine, under the optimal control policy for a given network configuration, how many pathways were allocated sufficient control to support task performance. Toward this end, we considered all pathways as active that were assigned a level of control of $k_p \geq .5$ for a given control policy, and denoted the number of such pathways as K. We selected .5 as the threshold for considering a pathway active on the basis of simulations using a wide range of parameter values and network configurations. For the I/O-matching model, this threshold corresponded to ~75 % match of its output to its input. For the DD model, it corresponded to harvesting ~72 % of maximal possible reward rate along a pathway. The number of active pathways K, defined in this way, provided an index of the limits on multitasking, with a smaller number indicating more severe limitation.

Network configuration

Using the procedures described above, we conducted simulations to examine how three factors determining network configuration impacted the limits on multitasking (as defined above): degree of pathway overlap (multiplexing), the extent to which this produced cross-talk (interference), and network size.

Pathway overlap (F) As we described under the Network Architecture section, each pair of input units projected to at least one pair of output units that comprised its pathway in the P_{rel} set. In addition, it could also project to other pairs of output units, constituting pathways in the P_{irrel} set. The number of such pathways was determined by a parameter F (for "fan-out"). This specified the number of output unit pairs to which pairs of input units projected (beyond the one in the P_{rel} set), and could range from 0 (no "fan-out" connections) to $N − 1$ (connected to all output unit pairs). Thus, F determined the amount of pathway overlap in the network, and consequently the amount of multiplexing. It is intuitively obvious that for $F = 0$ full multitasking is optimal, because there is no opportunity for cross-talk. What was of interest was how increasing F interacted with other network factors to influence the limits on multitasking. In determining irrelevant pathways, we simply assigned each input unit pair F random output unit pairs, making sure not to doubly connect the same input unit to an output unit. We also tried various other schemes for connecting inputs to outputs (such as nearest neighbor connections), all of which did not produce noticeably different results (see SOM B for details). In the analyses below, we considered the effects of pathway overlap both with respect to its absolute value (as described above), and with respect to its value as a proportion of network size, F/N.

Incongruence (φ) Increasing pathway overlap increased multiplexing, and therefore the potential for cross-talk in the network. In principle, such cross-talk could either enhance or degrade multitasking performance, depending upon whether information arriving at a pair of output units from the irrelevant pathway(s) was congruent with information arriving along the relevant pathway, and therefore facilitated processing, or was incongruent and therefore interfered with processing. This was determined by a parameter φ that specified the frequency with which processing along irrelevant pathways was congruent versus incongruent with processing along the relevant pathways.[6] This ranged from 0 (*full congruency*, i.e., activated inputs units in irrelevant pathways all had connection weights to the output units with the same sign as the activated unit in the relevant one) to 1 (*full incongruency*). Intuitively it is clear that for $\varphi = 0$ full multitasking will be optimal, whereas as for $\varphi = 1$ multitasking will be deleterious. What was of interest was how intermediate values of φ interacted with other network factors to influence the limits on multitasking.

Network size (N) Although it is clear that increasing pathway overlap together with incongruence in our models would limit multitasking, what was not clear was how restrictive this effect would be, nor how it would scale with network size (i.e., the number of relevant pathways). The latter is particularly important, given the large number of potential pathways in the human brain. For example, if the effects of pathway overlap and/or interference scaled linearly with network size (enforcing a limit of some fixed percentage of available pathways), this would provide a poor explanation for the remarkably constrained ability that is observed in humans for multitasking involving controlled-processing (i.e., even a small percentage of a *very* large number is also a large number). Rather, these effects would need to exhibit dramatically sublinear scaling to explain the striking disparity between the size of the human brain and its limited ability to multitask using controlled processing. To test for this, we simulated networks ranging in size from $N = 10$ to 1,000 pathways, evaluating the influence of pathway overlap and interference on the limits of multitasking at each scale.

---

[6] Specifically, processing was congruent for two overlapping pathways if the signs of the connection weights from the most highly activated input units in the two pathways to the shared output unit were the same. Since the convention, in both models, was for the left input unit in each pathway to receive greater activity than the one on the right, congruency could be manipulated by varying the frequency with which connections from the left and right input units in the irrelevant pathways had the same sign as (i.e., were "parallel" to) connections from the left and right input units in the relevant pathway, or conversely had the opposite sign (i.e., were "crossed"). This was controlled by φ. In Fig. 3, incongruence is illustrated by the connections of the irrelevant pathways (in a slightly lighter shade) having colors opposite those of the corresponding connections in the relevant pathways.

Results

Figure 4 shows the performance of the I/O-matching and DD models for different levels of interference φ and a value of pathway overlap of $F = 20\%$ (i.e., $F = N/5$). We only consider cases in which $\varphi > .5$, since this is the range in which the irrelevant pathways cause interference.

As expected, increasing incongruence increased the limits on multitasking, as is evidenced by a reduction in the number of active pathways K. Under maximal interference $\varphi = 1$, even as the network size increased well beyond 10 pathways, K converged to a value of around 10 in both models. In the DD model, asymptotic convergence seems to occur around a value of $\varphi = .75$. At that level of incongruence, even in the I/O-matching model the number of active control representations grew very slowly with network size, with only 17 active control representations in a network of size 100, and only three more with a tenfold increase in network size (to $N = 1,000$). On the basis of these observations, we use a fixed value of interference of $\varphi = .75$ in our subsequent simulations examining the effects of pathway overlap.

Interpreting the empirical significance of the interference parameter value is, of course, difficult. However, it does not seem unreasonable to assume, for performance of a task focused on a particular dimension of information, that information from other, irrelevant dimensions is distracting 75 % of the time. Indeed, it seems possible that this is an underestimate, given that our models focus on tasks involving only two choices; that is, in which the probability of incongruence by chance is 50 %. As the number of choices increases, the likelihood of incongruence by chance increases quickly. These considerations, coupled with the observation that this is the lowest value at which the effects of performance approach asymptote, suggest that $\varphi = .75$ is a reasonable (and perhaps conservative) point of reference.

Figure 5 shows how multitasking in the I/O-matching model is affected by pathway overlap (F) in networks of various sizes. The effects are shown both in terms of absolute values of F and K, and when these are expressed as fractions of network size N. As anticipated, increasing F quickly reduces K; that is, it increased the limits on multitasking. However, there are two important features of this effect: First, it is highly nonlinear, with even modest amounts of pathway overlap (e.g., a value of $F = 20$ %, upper right panel) producing a considerable reduction in the number of tasks that can be supported. Furthermore, this effect accelerates with network size. That is, there is substantial sublinearity, so that the strong constraint on the degree of multitasking—in terms of the absolute number of tasks that can be performed—is only minimally influenced by network size. For example, as mentioned above, even for a network with 1,000 pathways, pathway
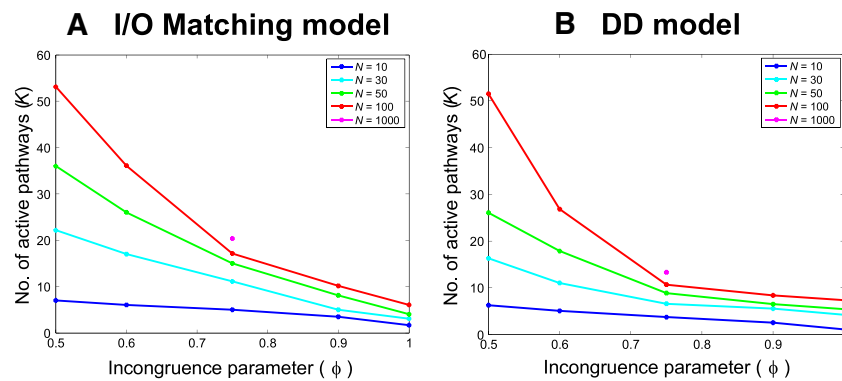
A I/O Matching model

B DD model

Fig. 4 Capacity constraints as a function of incongruence for networks of different sizes. The plots show the numbers of pathways active ($k > 0.5$; see the text for explanation) under the optimal control policy for different levels of incongruence (φ, along the x-axis) and networks of different sizes (N; different colored lines). All results are for a pathway overlap of $F = 20\%$ (i.e., $F = N/5$). Due to the computational demands of our model, we were only able to simulate networks of size 1,000 for $\varphi = .75$, $F = 20\%$. **(A)** I/O-matching model. Note the steeper slope of decline for larger networks, with convergence (at the highest levels of incongruence) to a value of about ten active pathways, irrespective of network size. Even at $\varphi = .75$ (the value used for most simulations), the number of pathways that can simultaneously be engaged increases only by a factor of 3 for an increase in network size of a factor of 100 ($K = 7$ for $N = 10$; $K = 23$ for $N = 1,000$). **(B)** DD model. The I/O-matching effects are substantially accentuated for the DD model, with near convergence for networks of all sizes at $\varphi = .75$

overlap of 20 % limits the amount of multitasking to below 30 pathways.

As with φ, it is difficult to relate F directly to a comparable quantity in the brain. However, a value of 20 % seems reasonable, if not modest, given the high degree of convergence for pathways in the brain. For example, using two publicly available large-scale cortical connectivity data sets (Rubinov & Sporns, 2010), compiled from tract-tracing studies in the macaque brain (in which nodes represent cortical areas and links represent large corticocortical tracts), yields convergence estimates of 15 % (using data from Young, 1993) and 23 % (using data from Sporns, Honey, & Kötter, 2007). It is also worth noting, again, that our analyses using the I/O-matching model provide what is likely to be a lower bound on the effect of pathway overlap and interference on multitasking. The results for the DD model support this conjecture.

Figure 6 shows results for the DD model analogous to those shown in Fig. 5 for the I/O-matching model. As a comparison of these results makes clear, the DD model exhibits effects similar to those of the I/O-matching model, only in accentuated form. For example, for network size $N = 1,000$ and $F = 20\%$, multitasking was limited to less than 14 active pathways, considerably below the level for the comparable network in the I/O-matching model. Figure 7 summarizes these effects, showing how the ability for multitasking changes with network size (for $F = 20\%$), with the DD model exhibiting a lower slope and approaching a substantially lower asymptotic value than the I/O-matching model. In the General Discussion, we consider factors that may explain these differences, and others that may further accentuate the effects of multiplexing on multitasking.

Several important questions could be asked about these results. A first question is: How sensitive was performance to the optimal value of K? For example, increasing the number of active control units beyond this value could be associated with only small costs to performance, calling into question the importance of this limit for multitasking. To examine this, we tested performance using values of K above and below the optimal value in both the I/O-matching and DD models. The results (shown in SOM D, Fig. S5) confirmed that, in both cases, maximum performance was achieved for the value of K identified as optimal (validating the optimization procedures), and that even small deviations from this value were associated with considerable decrements in performance. We also observed that increasing the number of active control units was more detrimental than decreasing that number, indicating that a conservative policy should favor limiting rather than licensing multitasking.

A second question is, How sensitive are these effects to the objective function used to determine the optimal control policy? This is perhaps less relevant to the DD model, for which the objective function—reward rate—both has face validity and has been used extensively in other theoretical work on the DD model (Balci et al., 2011; Bogacz et al., 2006; Simen, Cohen, & Holmes, 2006). Thus, we focused our attention on the I/O-matching model. In SOM D, we report results using a simpler objective function (sum of absolute errors; see Fig. S5, panel C) and show that these results do not differ qualitatively from those reported above.

Finally, it is reasonable to ask why the limits in multitasking that we observed were "hard"; that is, why scaling the optimal value of K with network size was so severely sublinear. It is difficult to give a precise answer to this question, since the high dimensionality of the system, coupled with the nonlinearities of the objective function, made it intractable to closed-form analysis (hence, the need for simulations). However, in SOM E, we report additional simulations using the I/O-matching model that provide some insight into the
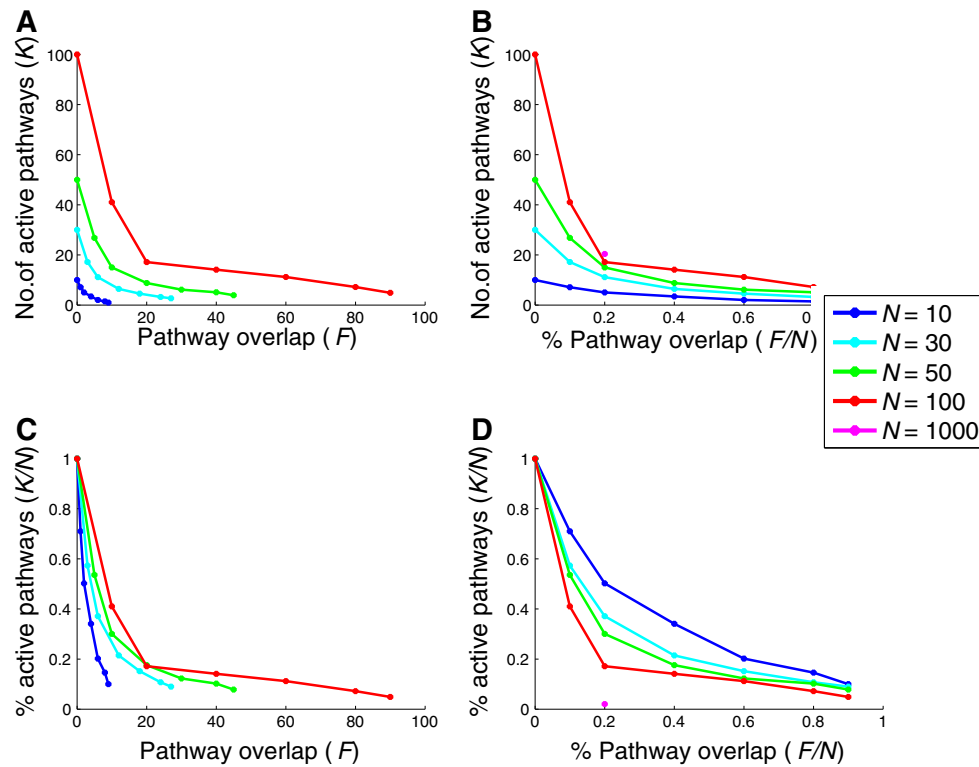
I/O Matching model

Fig. 5 Capacity constraints in the I/O-matching model as a function of pathway overlap for networks of different sizes. These results are for an incongruence of $\varphi = .75$. The top panels (**A** and **B**) show capacity constraints under the optimal control policy in terms of the absolute number of active pathways (K), and the bottom panels (**C** and **D**) show the results in terms of the percentage of active pathways (K/N). The left panels (**A** and **C**) show the results in terms of the absolute amount of pathway overlap (F), and the right panels (**B** and **D**) show them in terms of the percentage of pathway overlap (F/N). Note that when there is no overlap ($F = 0$), all relevant pathways are active ($K = N$), indicating full multitasking. Increasing F quickly drives this down, limiting multitasking. This effect is substantially greater for larger network sizes, with networks of all sizes converging to a similar limit in multitasking at around $F = 20\%$

source of these effects. They suggest that, whereas modulation of the entire pathway (i.e., both the input and output layers) produces optimal performance, the sublinear scaling of optimal K with network size relies heavily on control at the output layer (see SOM E, Figs. S6–S9). Some intuition for this effect can be gained by considering the extended version of the Stroop task described in the introduction. Presented with the options of performing both the color-naming and word–action tasks or just one of these, it would almost certainly be preferable to choose the latter. Assume, for the purposes of illustration, that the word–action task is chosen. In this case, the full model would modulate the activity of both the intermediate and response (output) units of the color-naming pathway, leaving only the word–action pathway activated (see Fig. 2B). However, now suppose that it were not possible to modulate the output units in a pathway. This would fail to suppress the effect of the intermediate units in the word pathway on the verbal response units, which (for incongruent stimuli) would produce an error in the color-naming task (since the inputs to those units from the intermediate units in the color pathway *were* suppressed). Therefore, under these

conditions, it would be preferable to also allocate control to the color-naming pathway, allowing its input and intermediate units to influence the response, and at least partially counteract the effects of the cross-talk at the output layer. In other words, in the absence of the ability to control the output, activating pathways that are subject to some (but perhaps not too much) cross-talk can contribute beneficially to their own outputs, and in some cases outweigh the cost of any additional cross-talk that they introduce into the network. However, when output can be controlled, this is not necessary, and the optimal policy is to limit processing to only the best-performing pathways.

General discussion

The simulations reported above provide a quantitative examination of the effects of pathway overlap on the number of tasks that can be simultaneously performed by networks with the capability of performing multiple two-alternative forced choice decision tasks. The results reveal that increasing
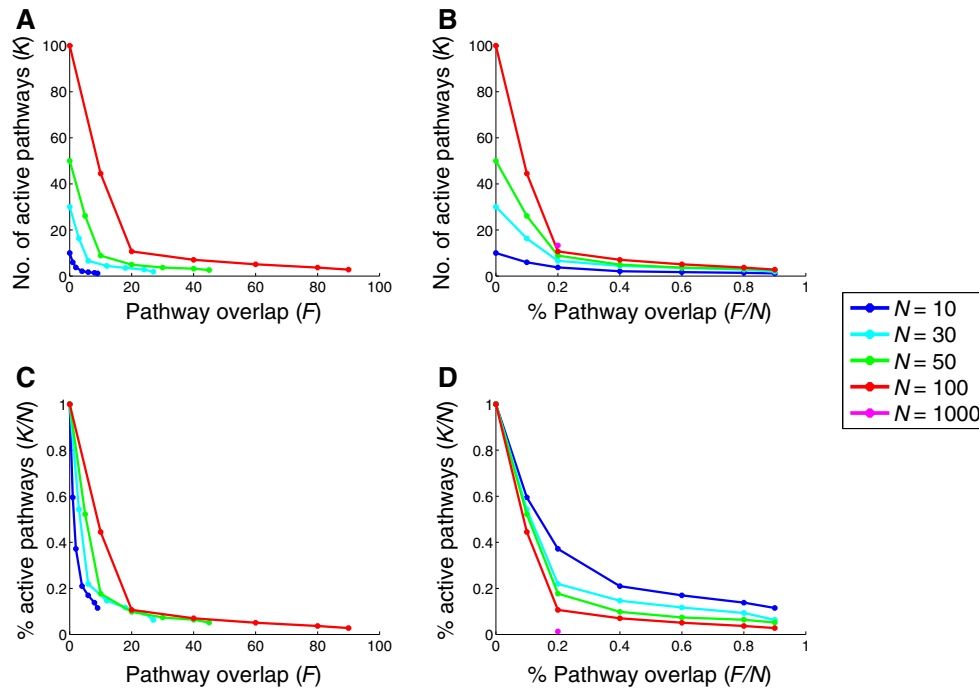
DD model

Fig. 6 Capacity constraints in the DD model as a function of pathway overlap for networks of different sizes. See the Fig. 5 caption for details. As in Fig. 4, the effects are substantially more pronounced than in the I/O-matching model, with a striking convergence in the limit on multitasking to less than 14 active pathways for a level of pathway overlap of $F = 20\%$, irrespective of network size

overlap rapidly constrains the number of tasks that can be performed at once, and that this reaches a maximum in a manner that is only weakly influenced by network size. These findings support the idea that multiplexing of representations in the brain may be an important source of limits in the ability for multitasking of control-dependent tasks. That said, the upper limits on multitasking observed in the models (10–30 tasks for a wide range of parameters) are noticeably higher than those typically observed for humans (certainly under 10). This may be due to a number of factors.

One factor is suggested by the difference in results observed for the I/O-matching and DD models. The limits on multitasking were less severe in the simpler, fully linear I/O-matching model than for the DD model. Two primary differences in the latter model were the addition of an integration process and the added nonlinearity that this introduced in the relationship between processing and performance. Presumably, the integration process afforded an additional opportunity for cross-talk from irrelevant pathways to interfere with processing in the relevant ones. The nonlinearity may also have contributed to this effect. The inclusion of additional nonlinearities commonly used in neural-network models (e.g., in the processing function itself) could be expected to further accentuate it (e.g., by amplifying the influence of activation in irrelevant pathways on relevant ones). Other factors that are likely to have a similar effect are multiple choices (increasing the likelihood of incongruence

among processes), recurrent connectivity (increasing interaction among processes), more complex processes involving
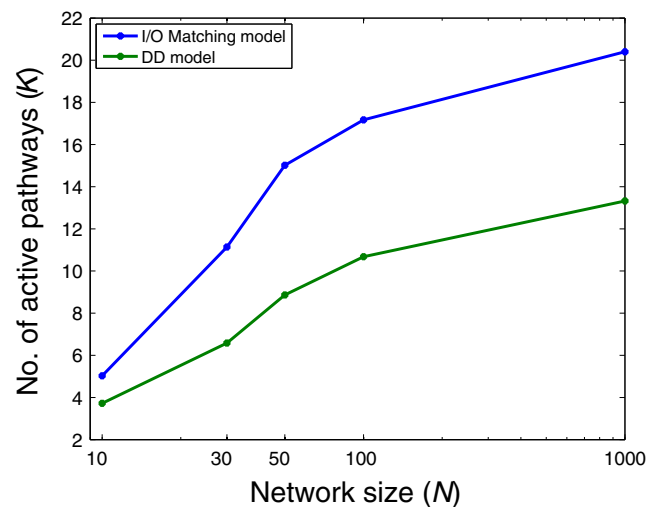

Fig. 7 Capacity constraints as a function of network size. The optimal control policy is plotted for networks of varying size ($N = 10$ to $1,000$) for the I/O-matching (*top curve*) and DD models (*bottom curve*). These results are for a pathway overlap of $F = 20\%$ and incongruence of $\varphi = .75$. Note that the slopes of both lines reduce considerably as network size increases, with the slope for the DD model decreasing more quickly and reaching a lower value than the slope for the I/O-matching model. See the text for a discussion. (Note that the change in slope from $N = 10$ to 30 is due to a disproportionately high number of active connections for small networks—about 50% of the units in the smallest network are active; this small effect vanishes as N exceeds 30.)

multiple levels of representation (increasing the opportunities for pathway overlap—such as the reading/dictation example in the introduction), asymmetric pathway strengths (possibly increasing reliance on control, as in the Stroop task), and distributed representations (contributing directly to multiplexing). None of these were implemented in the present models, but all of them are common features of real-world tasks and neural systems. The combined effect of these factors is likely to substantially potentiate the constraining effects of pathway overlap on the ability to carry out multiple simultaneous processes. Of course, additional work will be needed to determine whether and how these add to the constraints on multitasking. It is important to emphasize, however, that none of these factors alone would constrain multitasking in the absence of pathway overlap—that is, in the absence of multiplexing. In this respect, multiplexing can be viewed as a primary causal factor, with which these others may interact.

Our models suggest that multiplexing can impose strict limits on multitasking of control-dependent processes, without any recourse to limitations in the control system itself. That is, we imposed no *intrinsic* limitation on the number of control units that could be activated at once, nor did their engagement carry any direct costs or penalties—only those associated with the consequences on performance. Of course, introspection, traditional dogma (e.g., Posner & Snyder, 1975), and recent experimental findings (Dixon & Christoff, 2012; Kool & Botvinick, 2012; Kool, McGuire, Rosen, & Botvinick, 2010; Westbrook, Kester, & Braver, 2013) all suggest that the exertion of control *does* carry an intrinsic cost, in the form of "mental effort." An intriguing explanation for this may be that such costs, experienced as subjective disutility, reflect intrinsic biases limiting the engagement of control signals in order to minimize the risk of cross-task.

Although our work focused on the role of multiplexing in limiting multitasking, it does not preclude—and may complement—models that address potential constraints on the capacity of control mechanisms themselves. The existence of such constraints has consistently been suggested by one remarkably robust phenomenon—the psychological refractory period (PRP; Welford, 1952). This refers to the observation of a seemingly immutable cost associated with performing more than one control-demanding task at once, as compared to performing them each individually. The PRP has been offered as evidence that the capacity constraints of a centralized control mechanism impose a "central bottleneck" that prohibits truly concurrent multitasking, requiring instead that control-demanding tasks be carried out in sequence (e.g., Pashler, 1984). However, more recent accounts suggest that these effects may reflect the influence of task instructions rather than a central bottleneck (e.g., Howes, Lewis, & Vera, 2009; Schumacher et al., 2001). Nevertheless, several theories continue to propose that multitasking and control may rely on a centralized mechanism, and neuroscientific evidence has been martialed in support of this claim (Duncan & Owen, 2000; Roca et al., 2011; Tombu et al., 2011).

One influential class of theories has used production system architectures to develop computationally explicit models of multitasking performance. Traditionally, these have exploited a fundamental feature of this architecture—reliance on goal representations in declarative memory for the execution of control-demanding behavior—to explain constraints on multitasking (e.g., Byrne & Anderson, 2001). Using this framework, more recent threaded-cognition models have proposed that bottlenecks can exist both in the central mechanisms responsible for control (such as declarative memory) and in peripheral modules responsible for task-specific processes, and that constraints on multitasking can arise from either or both, depending on the specific characteristics of the tasks involved (e.g., Salvucci & Taatgen, 2008). However, even these models assume that the execution of all tasks relies on coordination by a common, serial procedural resource. Thus, whereas these models have produced remarkably detailed accounts of human multitasking performance in a variety of tasks, they risk stipulating (in the form of constraints on centralized mechanisms) the effects to be explained. At least one production system model, EPIC (Meyer & Keiras, 1997), occupies the far extreme of parallelism, accounting for constraints on multitasking performance under a variety of conditions entirely in terms of (i.e., as a response to) the conflicts that arise among shared peripheral processors.

Our models share an obvious similarity with the threaded-cognition model and, even more so, the EPIC model, as well as the early multiple-resource theories discussed in the introduction, in emphasizing the importance of cross-talk among interacting task-specific pathways as a constraint on multitasking. However, the focus of the production system models has been largely on the challenges posed by prioritizing and/or scheduling processes that are in potential conflict, and using these to account for the details of human multitasking behavior. Our work might be viewed as complementary to such work, addressing in a more general way a more proximal issue: the quantitative relationship between multiplexing, the demands for control that it poses, and the simultaneous limits that it seems to impose on the licensing of control. The simple and generic nature of our models suggests that our observations may represent general properties of systems involving interacting processes, which may be relevant to a broad class of models.

In a separate line of work, neurally inspired models of a very different flavor from the ones presented here have been proposed to explain capacity limits in working memory systems—that is, in the number of representations that can be actively maintained (e.g., Haarmann & Usher, 2001; Ma & Huang, 2009; Todd, Niv, & Cohen, 2009; see also Oberauer & Kliegl, 2006). These may be relevant to limits in the capacity for control-dependent processing, insofar as most models of controlled processing (whether neurally inspired or more

abstract) assume that this relies on the active maintenance of control signals, and therefore relies on some component of the working memory system (e.g., on task rules, intentions, and/or goals; e.g., Anderson, 1983; Cowan, 2001; Engel, 2002; E. K. Miller & Cohen, 2001). Thus, constraints on the capacity for active maintenance may also contribute to limits in the capacity for control-dependent processing, and integrating such models with the work presented here (e.g., introducing attractor dynamics into the control representations in our model) represents an interesting and potentially valuable direction for future research.

Even allowing for the possibility that capacity constraints are imposed by the characteristics of the control system itself, the present work provides support for the long-standing idea, suggested by multiple-resource theories, that the architecture of the rest of the processing system—over which the control system must preside—represents an important source of such constraints. Importantly, it brings this idea into contact with the growing efforts to consider control from a normative perspective (Botvinick & Cohen, in press; Dayan, 2007; Shenhav et al., 2013; Todd et al., 2009). A normative approach to theory may be particularly relevant to an understanding of controlled processing, which can usefully be defined as the set of mechanisms responsible for optimizing processing in the service of maximizing cumulative reward. One recent theory (EVC) proposes that, from this perspective, a critical function of the control system is to estimate the expected value of candidate control-demanding processes and specify control signals in a way that maximizes this quantity (Shenhav et al., 2013). This must take into account not only the expected benefits from controlled processes, but also the potential costs associated with their execution. Such costs include the potential for interference from cross-talk that arises from simultaneously engaging multiple processes. The work that we have presented here examined how such costs can figure into the choice of an optimal control policy, and in particular how it may favor a policy that restricts the allocation of control to a limited number of processes. Put simply, it makes "sense" not to try to do more things than can be done, since the expected returns for doing so will be low. From this perspective, our findings are consistent with the EVC theory, providing a normative account of why the control system may actively choose to limit multitasking.

An important assumption of this account is that multiplexing is either a predetermined feature of the processing architecture in the brain, or that its benefits outweigh the costs that they impose on multitasking. Multiplexing would seem to be a fixed characteristic of at least some parts of the processing system—for example, the motor system. That is, it seems reasonable to think that the limited dimensionality of the motor system, relative to the large space of potential stimuli to which we can respond, is one factor that imposes multiplexing (e.g., we don't have two mouths, one to respond

to the color and the other to the word in the Stroop task). However, this cannot account for the capacity constraints on control more generally. Internal processes (such as memory retrieval, mental imagery, planning, and problem solving) are not constrained by the dimensionality of the motor system, and yet clearly exhibit the same strict limitations on multitasking (i.e., a "single line of thought) as do more overt control-demanding behaviors. This suggests that if limitations of "internal multitasking"—like those of overt behaviors—reflect constraints imposed by multiplexing, then there must be some intrinsic benefits to the multiplexing of the internal representations that support such processes. We suggested some of these in the introduction, such as the efficiency of coding and inference (Forbus et al., 1995; Hinton et al., 1986). This may also include learning (e.g., the use of distributed, similarity-based representations to reduce the state space for reinforcement-learning mechanisms).[7] It will be important, in future work, to explore the relationship between multiplexing and multitasking, by identifying meaningful objective functions that can be used to evaluate the worth of multiplexing representations, and by integrating these with the ones for processing (used to evaluate multitasking) that we have explored here.

Although more comprehensive normative analysis is clearly an important goal, the findings reported here provide a strong hint concerning their likely outcome. The interaction between multiplexing and multitasking that we observed was heavily asymmetric: Even modest degrees of multiplexing quickly and strongly diminished the value of multitasking. This suggests that, in a trade-off between the two, only modest benefits to multiplexing need be present to normatively justify strong constraints on multitasking.

Finally, it is important to consider whether and how the work we have presented here aligns with work addressing related questions in other fields. Concerns about cross-talk and process scheduling are fundamental to many areas of engineering (such as computer operating-system design, network communications, and transit systems) and other areas of the life sciences (e.g., genetic expression and cell migration during development). Accordingly, these issues have prompted the development of mathematical tools in control theory, network topology, and graph-theoretical analysis. However, these tools and their application have focused either on optimizing the scheduling of sequential processes or increasing and maintaining the robustness of a complex system (Carlson & Doyle, 2002). To our knowledge, no theoretical or applied work has specifically addressed the intersection of factors relevant to the problem that we have outlined in this article: the constraints on parallel processing introduced by

---

[7] We thank Mike Mozer for pointing out this possibility, as well as the potential link that it provides between the account of capacity constraints explored here and the one concerning the role of working memory in reinforcement learning proposed by Todd et al. (2009).

cross-talk from the multiplexing of representations upon which those processes depend. Because of the nonlinearity and high dimensionality of the optimization problem, we were forced to resort to numerical computations. Ideally, we would like more rigorous, analytic tools for quantifying the constraints that multiplexing imposes on multitasking within the processing architecture of neural systems. In this respect, the problem that we have posed represents a novel and potentially interesting future challenge for mathematical analysis.

In summary, we have presented simulations that outline the constraints on multitasking imposed by the multiplexing of representations. The latter demands the engagement of control mechanisms in order to prevent cross-talk. From this perspective, the very feature of the system that demands control, and yet contributes to its flexibility—multiplexing—appears at the same time to limit the number of control-demanding tasks that can be executed at once. As we discussed above, this is likely to interact with other features of controlled processing, which together may explain the remarkable limits observed in the human ability for multitasking. To date, research efforts in this area have reflected a level of parallel multitasking that appears unbound by any constraints on control. Hopefully this has begun to produce insights that, with more interactive efforts in the future, will lead to a more complete and coherent understanding of limitations in the performance of control-demanding behaviors in humans—behaviors that are otherwise some of the most flexible, powerful, and characteristic of those exhibited by humans.

References

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

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,* 225–235.

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

Baddeley, A. D. (1970). Estimating the short-term component in free recall. *British Journal of Psychology, 61,* 13–15.

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

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. doi:10.3758/s13414-010-0049-7

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

Berlyne, D. E. (1957). Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts. *Psychological Review, 64,* 329–339. doi:10.1037/h0041135

Birgin, E. G., & Martínez, J. M. (2008). Improving ultimate convergence of an augmented Lagrangian method. *Optimization Methods and Software, 23,* 177–195.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113,* 700–765. doi:10.1037/0033-295X.113.4.700

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

Botvinick, M., & Cohen, J. D. (in press). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science.*

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.), *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737). Cambridge: MIT Press.

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks, 17,* 471–510.

Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychological Review, 108,* 847–869. doi:10.1037/0033-295X.108.4.847

Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences, 99*(Suppl. 1), 2538–2545.

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

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 71–90). New York: Guilford Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361. doi:10.1037/0033-295X.97.3.332

Conn, A. R., Gould, N. I., & Toint, P. (1991). A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis, 28,* 545–572.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi:10.1017/S0140525X01003922. disc. 114–185.

Dayan, P. (2007). Bilinearity, rules, and prefrontal cortex. *Frontiers in Computational Neuroscience, 1*(1), 1–14. doi:10.3389/neuro.10.001.2007

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91,* 176–180.

Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: Neural and behavioral evidence. *PLoS ONE, 7,* e51637. doi:10.1371/journal.pone.0051637

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23,* 475–483.

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the Sixth International Symposium on Micro Machines and Human Science, 1995 (MHS '95)* (pp. 39–43). Piscataway, NJ: IEEE Press

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

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

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2,* 493–501.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience, 10,* 1563–1573.

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

Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review, 8,* 568–578.

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

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience, 3,* 31.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing. Volume 1: Foundations.* Cambridge: MIT Press.

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

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,* 717–751. doi:10.1037/a0017187

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17,* 187–201. doi:10.1017/S0140525X00034026. disc. 201–245.

Johnson, S. G. (2012). The NLopt nonlinear-optimization package (Version 2.3). Retrieved from http://ab-initio.mit.edu/nlopt

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122–149. doi:10.1037/0033-295X.99.1.122

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.

Kan, A. R., & Timmer, G. T. (1987). Stochastic global optimization methods part I: Clustering methods. *Mathematical Programming, 39,* 27–56.

Kool, W., & Botvinick, M. (2012). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General.* doi:10.1037/a0031048. Advance online publication.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of*

Experimental Psychology: General, 139, 665–682. doi:10.1037/a0020198

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

Kriete, T., & Noelle, D. C. (2011). Generalisation benefits of output gating in a model of prefrontal cortex. *Connection Science, 23,* 119–129.

Liu, Y. S., Holmes, P., & Cohen, J. D. (2008). A neural network model of the Eriksen task: Reduction, analysis, and data fitting. *Neural Computation, 20,* 345–373.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology, 39,* 367–386. doi:10.1037/h0080066

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9*(11:3), 1–30. doi:10.1167/9.11.3

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203. doi:10.1037/0033-2909.109.2.163

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,* 749–791. doi:10.1037/0033-295X.104.4.749

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

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

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology, 74,* 774–789.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review, 86,* 214–255. doi:10.1037/0033-295X.86.3.214

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18,* 283–328. doi:10.1162/089976606775093909

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex, 12,* 246–257.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language, 55,* 601–626. doi:10.1016/j.jml.2006.08.009

Pakkenberg, B., & Gundersen, H. J. (1997). Neocortical neuron number in humans: Effect of sex and age. *Journal of Comparative Neurology, 384,* 312–320.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 358–377. doi:10.1037/0096-1523.10.3.358

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale: Erlbaum.

Powell, M. J. D. (2006). *The NEWUOA software for unconstrained optimization without derivatives. Large-scale nonlinear optimization* (pp. 255–297). New York: Springer.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives (Technical Report No. DAMTP 2009/NA06).* Cambridge: University of Cambridge, Department of

Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. doi:10.1111/1467-9280.00067

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300. doi:10.1037/0033-295X.106.2.261

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 flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences, 102,* 7338–7343.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage, 52,* 1059–1069.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115,* 101–130. doi:10.1037/0033-295X.115.1.101

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience, 2,* 33–42.

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.

Servan-Schreiber, D., Bruno, R. M., Carter, C. S., & Cohen, J. D. (1998). Dopamine and the mechanisms of cognition: Part I. A neural network model predicting dopamine effects on selective attention. *Biological Psychiatry, 43,* 713–722.

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

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

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,* 127–190. doi:10.1037/0033-295X.84.2.127

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks, 19,* 1013–1026.

Simen, P., Contreras, 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.

Sporns, O., Honey, C. J., & Kötter, R. (2007). Identification and classification of hubs in brain networks. *PLoS ONE, 2,* e1049. doi:10.1371/journal.pone.0001049

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654. doi:10.1126/science.153.3736.652

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662. doi:10.1037/0096-3445.121.1.15

Thune, J. J., Uylings, H. B., & Pakkenberg, B. (2001). No deficit in total number of neurons in the prefrontal cortex in schizophrenics. *Journal of Psychiatric Research, 35,* 15–21.

Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In D. Koller (Ed.), *Advances in neural information processing systems 21* (pp. 1700–1707). Red Hook: Curran Associates.

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, 108,* 13426–13431.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592. doi:10.1037/0033-295X.111.3.757

Vaz, A. I. F., & Vicente, L. N. (2009). PSwarm: A hybrid solver for linearly constrained global derivative-free optimization. *Optimization Methods and Software, 24,* 669–685.

Welford, A. T. (1952). The "psychological refractory period" and the timing of high-speed performance—A review and a theory. *British Journal of Psychology, 43,* 2–19.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS ONE, 8,* e68210. doi:10.1371/journal.pone.0068210

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.

Williams, R. W., & Herrup, K. (1988). The control of neuron number. *Annual Review of Neuroscience, 11,* 423–453.

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,* 931–959. doi:10.1037/0033-295X.111.4.931

Young, M. P. (1993). The organization of neural systems in the primate cerebral cortex. *Proceedings of the Royal Society B, 252,* 13–18.

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