Basic functional trade-offs in cognition: An integrative framework

Basic Functional Trade-offs in Cognition: An Integrative Framework

Marco Del Giudice

Bernard J. Crespi

***Cognition, 179,* 56-70 (2018).**

Marco Del Giudice, Department of Psychology, University of New Mexico;
Bernard J. Crespi, Department of Biological Sciences, Simon Fraser University.
Address correspondence to Marco Del Giudice, Department of Psychology, University of New Mexico. Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

Abstract

Trade-offs between advantageous but conflicting properties (e.g., speed vs. accuracy) are ubiquitous in cognition, but the relevant literature is conceptually fragmented, scattered across disciplines, and has not been organized in a coherent framework. This paper takes an initial step toward a general theory of cognitive trade-offs by examining four key properties of goal-directed systems: performance, efficiency, robustness, and flexibility. These properties define a number of basic functional trade-offs that can be used to map the abstract "design space" of natural and artificial cognitive systems. Basic functional trade-offs provide a shared vocabulary to describe a variety of specific trade-offs including speed vs. accuracy, generalist vs. specialist, exploration vs. exploitation, and many others. By linking specific features of cognitive functioning to general properties such as robustness and efficiency, it becomes possible to harness some powerful insights from systems engineering and systems biology to suggest useful generalizations, point to under-explored but potentially important trade-offs, and prompt novel hypotheses and connections between disparate areas of research.

Keywords: Design; efficiency; flexibility; performance; robustness; trade-offs.

1. Introduction

Trade-offs—balances between separately advantageous but conflicting traits—are fundamental aspects of all systems, whether they are artificial machines or biological mechanisms designed through evolution by natural selection. Trade-offs are also ubiquitous in cognitive systems. Enhanced computational performance does not come for free; the same is true of other desirable properties such as speed, flexibility, or the ability to withstand damage. Crucially, improving a system on one front will typically worsen it in other ways. For example, the speed of decisions can be increased by sacrificing their accuracy (Heitz, 2014), and more flexible learning algorithms also tend to be more computationally demanding (Daw & Dayan, 2014). The design of cognitive systems is thus shaped by constraints, compromises, and opposing priorities that can be understood only in relation to the underlying trade-offs.

Cognitive trade-offs have been addressed in many disciplines, from neuroscience and psychology to behavioral ecology and computer science. Unfortunately, the relevant literature remains scattered, limited in scope, and conceptually fragmented. Different research traditions tend to focus on different trade-offs, largely ignore each other's contribution, and often employ different terms for similar or overlapping constructs. To the best of our knowledge, there have been no attempts to organize this literature within a coherent framework. Here we take an initial step in this direction by offering an integrative overview of what we label *basic functional trade-offs*: a set of highly general trade-offs that apply to all natural or artificial systems designed to perform a function, including cognitive systems whose function can be described as manipulation of information (Piccinini & Scarantino, 2011; more on this in section 2).

Basic functional trade-offs are defined by four key properties of goal-directed systems: *performance*, *efficiency*, *robustness*, and *flexibility* (Figure 1). Together, these properties map the abstract "design space" of any natural or artificial system endowed with a function; when they are applied to cognitive systems (as we do here), they provide a shared vocabulary to describe a variety of specific characteristics such as speed, accuracy, reliability, memory use, and so on. By linking specific features of cognitive functioning to general properties such as robustness and efficiency, it becomes possible to harness some powerful insights from systems engineering and systems biology, two related disciplines that explicitly investigate the design of complex functional mechanisms (Alderson & Doyle, 2010; Doyle & Csete, 2011; Kitano, 2004, 2007).

We have identified the four properties in Figure 1 as basic after surveying an extensive literature on trade-offs in biology and engineering, as detailed in the remainder of this paper. We could not find other examples of properties that were both universal (i.e., would apply to all functional systems) and similarly general (i.e., were not already encompassed by the basic ones). This assertion does not mean that the classification we propose is fully exhaustive or that it cannot be extended in principle, and we encourage its growth and elaboration. As we discuss in the following sections, even the distinctions between basic properties are not absolute, and admit a degree of conceptual overlap—for example, in particular cases it can be hard to differentiate sharply between robustness and flexibility, or between robustness and performance. While it is important to acknowledge and discuss those cases, the functional properties that we describe have a broad range of application and considerable heuristic power. Their value lies in their

ability to integrate many particular examples within a common frame of reference, suggest useful generalizations, and prompt novel hypotheses and connections across scientific domains.

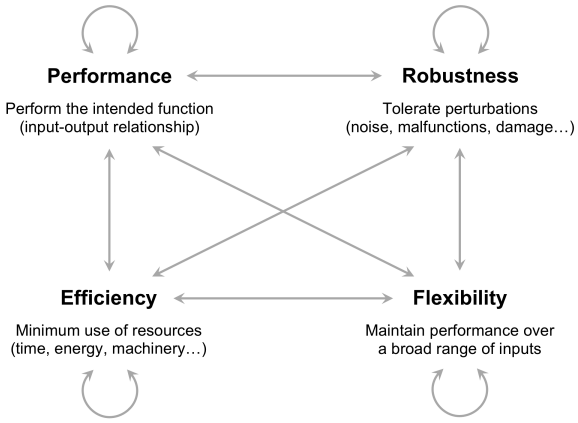

Figure 1. A map of basic functional trade-offs. Performance, efficiency, robustness, and flexibility are the key properties of all functional systems, including natural and artificial cognitive systems. Straight arrows represent trade-offs between properties; curved arrows represent trade-offs between different aspects of the same property.

In this paper, we examine the four properties shown in Figure 1 and discuss the trade-offs that arise between competing properties, as well as between different aspects of each (e.g., trade-offs between multiple aspects of robustness), with a focus on cognitive and neural systems. We also consider the implications of simultaneous trade-offs among more than two properties (e.g., three-way trade-offs between performance, robustness, and efficiency). The framework we present brings together many specific trade-offs that have been investigated in the literature (summarized in Table 1), points to some potentially important trade-offs that have received comparatively little attention so far, and offers a toolkit for clarifying some counterintuitive phenomena such as "less-is-more" effects in the performance of simple cognitive heuristics (Gigerenzer & Brighton, 2009). We conclude by considering possible ways to apply and extend the framework. From a psychological perspective, a better understanding of trade-offs may illuminate typical human cognitive variation as well as mental disorders, some of which appear to involve extremes or dysregulation in the balances between competing cognitive functions (e.g., Baron-Cohen, 2009; Crespi & Go, 2015).

2. Performance

The performance of a system is usually defined as its ability to produce an intended result (or some other roughly equivalent formulation). The concept of performance is meaningless without explicit or implicit reference to *function*, the idea that the system has an identifiable purpose, goal, or rationale. In turn, function implies *design*—in order to fulfill a purpose, a system needs to be structured in an organized, non-random fashion. When the term "design" is employed in this broad sense it does not require the existence of a conscious designer: indeed, the crucial insight of Darwinian biology is that design and function can arise from the blind, undirected, and impersonal process of natural selection (Alderson & Doyle, 2010; Dennett, 2009; Sterling & Laughlin, 2015).

Basic trade-offs	Main examples discussed in the text
Efficiency vs. performance	- Speed-accuracy trade-offs - Exploration-exploitation trade-offs - Efficiency trade-offs in neural design
Efficiency vs. robustness	- Robustness-resource trade-offs - Proactive vs. reactive control
Efficiency vs. flexibility	- Generalist-specialist trade-offs - Model-based vs. model-free learning - Fast and frugal heuristics
Performance vs. robustness	- Bias-variance trade-offs - Pessimistic strategies
Performance vs. flexibility	- Generalist-specialist trade-offs
Robustness vs. flexibility	- Stability-flexibility dilemma - Proactive vs. reactive control - Fast and frugal heuristics
Aspects of efficiency	- Space-time trade-offs
Aspects of robustness	- Robustness-fragility trade-offs

Table 1. Summary of the trade-offs discussed in the text.

In biological systems, goals can exist on an objective level even if they are not represented consciously (or at all) within the system. When bacteria move toward higher concentrations of glucose by chemotaxis, their behavior is regulated by a system of feedback control that alternates straight line swimming and random tumbling. The objective goal of this behavior is to move bacteria toward glucose, even if bacteria themselves have no representation of it—and, interestingly, do not even possess a representation of the direction in which they are swimming (Bechhoefer, 2005). Such real but unrepresented goals are ubiquitous in biological systems; in Daniel Dennett's terminology, they can be described as "free-floating rationales" produced by blind selection (Dennett, 2009). The difference between free-floating rationales and deliberate, fully represented goals (such as those of a human designer) is best understood as a gradient, which is climbed by evolutionary processes through the gradual accumulation of functional specialization and cognitive complexity. For the purpose of this paper, we make no distinction between different types of goals, and the concepts of design and function apply to natural and artificial systems alike.

2.1. Performance in Cognitive Systems

Broadly defined, a cognitive system is an information-processing mechanism that computes mappings between inputs and outputs (Lewis et al., 2014). Input-output mappings can be extremely complex; as well, outputs can take many possible forms, including commands to physical effectors (e.g., muscles or motors) but also representations that are used as inputs to other systems (e. g., information transfer between different brain regions). Note that we employ both "computation" and "information" in a broad sense, to include non-algorithmic and non-digital types of computation as well as various types of information (e.g., Shannon vs. semantic information; see Piccinini & Scarantino, 2011). Thus, our working definition of a cognitive system includes both natural and artificial instances of information-processing mechanisms. Some proponents of dynamical approaches to cognition (most notably van Gelder, 1998) have argued that cognitive mechanisms should be understood as dynamical systems—as defined for example by sets of differential equations—rather than computational processes. However, dynamical systems can also be analyzed with the tools of information theory and described from a computational perspective (Beer & Williams, 2015; Quax et al., 2016), making the distinction irrelevant for our purposes.

When the function of a system involves information processing, its performance can be assessed with respect to the intended relationships between input and output (Gluck et al., 2012). A system performs well when it produces the intended output in relation to a certain input; performance degrades to the extent that the actual output diverges from the intended one. What counts as "intended" depends on the specific function of the system, whether that function is explicitly represented in the mind of human designers (e.g., an application that detects and recognizes faces in pictures) or is a free-floating rationale in a biological mechanism (e.g., the face recognition circuits found in primate brains). Importantly, the intended input-output relationships can be probabilistic rather than deterministic; for example, a system may be designed to yield certain patterns of correlation between features of the input and those of the output. Moreover, we make no assumption that the output is solely or uniquely determined by the input; in particular, a cognitive system may produce self-generated output patterns that are at least partly independent from inputs.

Parsing the functions of complex cognitive systems can be challenging, and a detailed taxonomy of cognitive functions is outside the scope of this paper. Typical functions discussed in the literature include perception, detection, decision-making, memory, and motor control. These categories are not clearly demarcated and overlap substantially with one another; for example, perceptual processes can be partly understood as detection tasks, and may be deeply intertwined with motor control (e.g., eye saccades play an important role in visual perception). Information processing is also crucially involved in motivation, emotion, communication, and in the regulation of visceral and endocrine processes. Cognitive systems are often arranged hierarchically, with smaller/simpler systems nested within larger/more complex ones (e.g., hierarchies of routines and subroutines in software applications; hierarchies of neural circuits in brains, down to the level of individual neurons and synapses). Thus, what constitutes the system of interest—and its corresponding functions, input, and output—critically depends on the level of analysis one decides to adopt. Of course, it is only possible to speak meaningfully of

performance when the design logic of a system has been "carved at its joints" and its function has been correctly identified.

3. Efficiency

The efficiency of a system is its ability to perform its function with minimal use of resources. Time is a vital resource, particularly in cognition: a faster system can respond more quickly to important events, make rapid decisions, and free up time for other tasks. When the activity of a system relies on the serial (as opposed to parallel) concatenation of multiple subsystems, the delays introduced by each of them will cumulate, making speed a highly desirable property. Note that, by treating time as a resource, the present framework draws a distinction between computational performance (the ability to produce the intended input-output mappings) and computation speed (the time employed to produce those mappings). The distinction may become blurred in cases in which processing speed is an integral component of performance. For example, the ability to accurately process high-pitched acoustic signals critically depends on the bandwidth of neural transmission. Accordingly, auditory neurons spike at faster rates than do visual or olfactory ones, and have thicker and more expensive axons (Sterling & Laughlin, 2015). Since higher frequencies require proportionally higher spiking rates, this is an interesting case in which a system's performance—that is, the ability to encode and process high acoustic frequencies—cannot be separated from the speed of its components.

Another crucial type of resource is energy, which is required for the operation of any natural or artificial system. Computation can be quite energy-intensive, and selection to minimize energetic demands seems to have shaped the evolution of neural machinery, from the biophysical characteristics of neurons to patterns of brain connectivity (e.g., Hasenstaub et al., 2010; Lennie, 2003; Tomasi et al., 2013). The amount of physical machinery devoted to computation—processors, neurons, connections—is yet another aspect of efficiency. Each additional component of a cognitive system requires additional energy to build, maintain, and operate. Not least, larger and heavier systems are harder to transport, which is a critical limitation for animal brains, but also for computers that need to be embedded in small, portable, or moving devices. Selection for efficiency may contribute to explaining why modularity is a widespread property of biological networks, from metabolic pathways to brains. Modular networks are clustered—their elements are densely connected within each cluster but only sparsely connected with elements outside the cluster. Intuitively, modularity can enhance performance, as each module becomes more specialized for a particular task (Rueffler et al., 2012). Less obviously, evolutionary simulations show that selection to minimize the *cost* of the connections between the elements of a network (including the necessary machinery and the energy required to run it) may be sufficient to favor the evolution of a modular organization (Clune et al., 2013). Finally, complex cognitive systems often contain centralized computational resources (e.g. shared memory spaces) that can be accessed and used by multiple subsystems. From the perspective of each of the subsystems, efficiency includes minimal use of limited shared resources such as memory and attention.

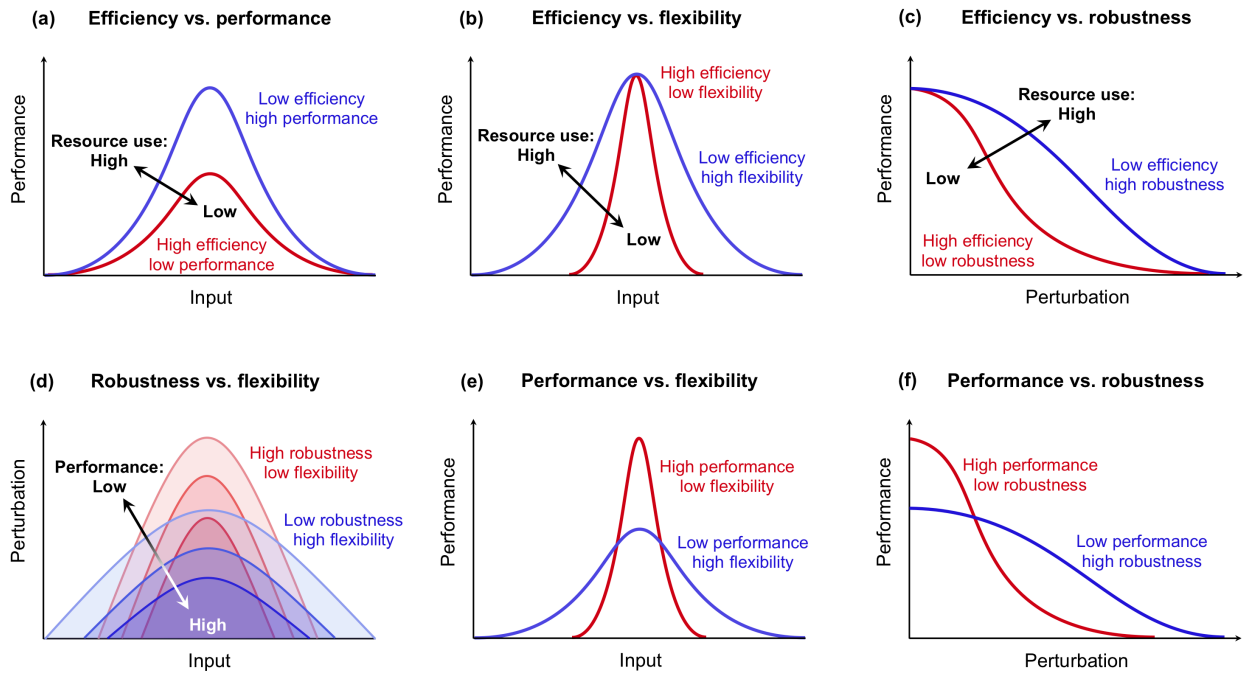

Figure 2. Schematic representations of pairwise trade-offs between performance, efficiency, robustness, and flexibility. Curves in panels (a), (b), and (e) represent the performance of a system in response to a range of inputs (horizontal axis); performance is maximal for a certain kind of input and degrades for increasingly different inputs. Taller curves indicate higher maximum performance; broader curves indicate higher flexibility. Curves in panels (c) and (f) represent the performance of a system exposed to increasingly severe perturbations (horizontal axis). Curves with a steeper downward slope indicate lower robustness. The contour plot in panel (d) represents the performance of a system (darker colors = higher performance) in response to a range of inputs (horizontal axis) and increasingly severe perturbations (vertical axis). The more flexible system (blue) maintains performance in response to a broader range of inputs, but its performance degrades more steeply as perturbations increase.

3.1. Trade-Offs Between Efficiency and Performance

The idea that reducing resource use may limit performance is an intuitive one, and there is a rich literature on trade-offs between efficiency and cognitive performance (Figure 2a). The best-known case is arguably that of *speed-accuracy trade-offs*, a broad class of phenomena in which faster performance on a task (i.e., more efficient use of time) leads to less accurate responding (Garrett, 1922; Heitz, 2014). In this context, "accuracy" refers to the ability to produce the correct answer to the task—for example detect or identify a target stimulus, make the correct decision, or find the correct answer to a problem. This usage of the term is commonplace in cognitive science but may cause some confusion with the way "accuracy" has been classically used in measurement theory to quantify systematic error (the distance between the true value and the mean of a set of measures), as contrasted with the random error quantified by *precision*. However, more recent metrological standards recommend using "trueness" to describe the lack of systematic error, and employ "accuracy" in the more ordinary sense of the overall closeness of a measurement to the true value (Joint Committee for Guides in Metrology,

2008). This revised meaning of accuracy is conceptually closer to the one found in the cognitive literature.

Speed-accuracy trade-offs have been documented in many different species and across a variety of tasks, from visual discrimination and motor control to predator avoidance and decision-making in foraging (Chittka et al., 2009; Sih & Del Giudice, 2012; Soukoreff & MacKenzie, 2009). In many contemporary cognitive models of decision-making (e.g., drift-diffusion models), the trade-off arises from the sequential sampling of information. By the logic of these models, more time spent sampling a noisy input translates into a better estimate of the actual state of the world, which in turn permits more accurate decisions (see Heitz, 2014). A similar trade-off applies to inference algorithms that progressively refine their estimates through multiple iterations (e.g., Bayesian sampling); in this case, each additional iteration improves accuracy but increases the time spent computing (Lieder et al., 2012). From yet another perspective, the existence of speed-accuracy trade-offs across domains may be predicted from some abstract properties of information transfer in noisy cognitive systems with limited capacity. Specifically, error rates can be expected to increase as the rate of information generated in the system exceeds the capacity of the transmission channel between input and output (e.g., the sequence of sensory, cognitive, and motor processes that determine the response to a stimulus in a discrimination task). When the rate of information transfer exceeds the channel capacity, every additional increase in speed reduces accuracy by increasing the error rate (Soukoreff & MacKenzie, 2009; note that this is a theoretical argument rather than an empirical generalization).

Another class of pervasive trade-offs involving time efficiency is that of *exploration-exploitation trade-offs* (Hills et al., 2015; Mehlhorn et al., 2015). In many types of cognitive tasks, it is possible to increase performance by sampling the environment for information and looking for additional options (exploration) instead of simply choosing the best option among those already known (exploitation). The benefits of exploration are typically uncertain, in contrast with the predictable outcomes of exploitation. The main cost of exploration lies in the time spent gathering information, although energetic costs may also play a significant role in some contexts. Finding the optimal balance of exploration and exploitation is a fundamental problem in the design of control systems and learning algorithms (including reinforcement learning; Dayan & Daw, 2008). Classic examples of this trade-off have been described in foraging (exploiting a known patch of food vs. searching for more abundant patches; Charnov, 1976) and mate choice (mating with a current attractive partner vs. waiting for better opportunities; Todd & Miller, 1999). Beyond decision-making, exploration-exploitation trade-offs occur in many cognitive processes that involve sequential searches under uncertainty—including visual attention, memory search, problem-solving strategies, and executive functions that regulate the allocation of cognitive resources between multiple tasks and goals (see Hills et al., 2015).

Efficiency constraints on cognitive performance are by no means limited to time. Several striking examples come from research in neural biophysics. Neurons are subject to a strong trade-off between the accuracy of information transmission through the synapse and the energetic cost of postsynaptic excitatory currents. Larger currents (which transmit information more accurately) can be obtained by increasing the influx of ions through postsynaptic channels;

however, the energy required to pump ions out and repolarize the membrane increases as well. The evidence indicates that synapses in visual pathways are not designed to maximize information transfer, but to optimize the *ratio* of information transmitted to energy consumed (Attwell & Laughlin, 2001; Harris et al., 2015; Sterling & Laughlin, 2015).

Other energetic trade-offs arise in patterns of functional connectivity in the brain: long-distance connections can increase computational performance but are considerably less energy-efficient than short-distance ones (Tomasi et al., 2013). The performance of neural transmission (in the form of enhanced signal-to-noise ratio) can also be improved by increasing the number of redundant synapses and/or their size, which drives up energetic consumption and requires additional cellular machinery. Accordingly, larger synapses are selectively expressed by neurons that perform high-precision computations (Laughlin, 2001; Sterling & Laughlin, 2015). Similarly, axons with a larger diameter and/or thicker myelination can transmit information faster and with higher signal-to-noise ratios. At the same time, they require larger amounts of lipids and proteins to build and take up more space, resulting in increased white matter volume and higher energetic consumption for building and maintenance (Wang et al., 2008). The design of the brain reflects myriad efficiency trade-offs at all levels of analysis, from the large-scale features of cortical structures and connections to the characteristics of single neurons and synapses (for extended discussion see Sterling & Laughlin, 2015).

3.2. Trade-Offs Between Aspects of Efficiency

From the preceding sections, it is easy to see how different aspects of efficiency may partly conflict with one another. As a rule, faster processors also tend to consume more energy; the same trade-off between energy and time efficiency applies to neurons, whose capacity for rapid spiking can be increased only by increasing the energetic cost of each action potential (Hasenstaub et al., 2010; Laughlin, 2001). Note that enhancing the response speed of individual neurons increases their ability to transmit information at faster rates, and thus encode signals with a higher bandwidth; this may translate into both speed *and* performance gains at the level of larger neural circuits. Thus, energy efficiency in neural transmission may trade off simultaneously against speed and performance.

Trade-offs between speed and energy consumption are relatively intuitive. Less intuitively, research in computer science has demonstrated the pervasive existence of *space-time trade-offs* in computation (Savage, 2008). The general idea is that the time required to perform a certain computation can often be reduced by increasing the memory space available to run the program. In cognitive science, space-time trade-offs have been discussed mainly in relation to the design of cognitive architectures (computational models of general-purpose intelligent agents; see Langley et al., 2009). For example, storing the results of previous computations into "chunks" of declarative knowledge or rules for action permits faster decisions, at the expense of memory space (e.g., Kurup & Lebiere, 2012; Rosenbloom et al., 1991). However, the concept has not been applied as broadly as it deserves in psychology and neuroscience. Trade-offs between memory space and processing time may partly explain the robust positive correlations between working memory capacity, processing speed, and performance on intelligence tests (which typically reflects response speed as well as accuracy; see Ackerman et al., 2005; Conway et al., 2002). At the neural level, space-time trade-offs may offer insights into the mechanisms of

cortical plasticity during skills acquisition. For example, intensive skill straining—which usually includes training for fast execution—is accompanied by expansion of the cortical regions involved in the task during the initial phase of performance gain (cortical representation may "renormalize" after performance stabilizes; see Ungerleider et al., 2002; Wenger et al., 2017; Zatorre et al., 2012). To the extent that larger cortical representation entails increased memory availability, it is possible that the recruitment of additional neural space is driven—at least in part—by requirements for increased computation speed while the new skills are being fine-tuned.

3.3. From Functional Trade-Offs to Design Solutions

The existence of functional trade-offs may shape the design of cognitive systems in a number of different ways. It can be useful to think about the consequences of trade-offs in terms of "design solutions" that balance competing properties within a certain set of constraints, which determine the range of feasible solutions. To begin, trade-offs may shape the basic structure and functionality of a given system. The biophysical properties of neurons and the general patterns of connectivity in the brain are case in points (e.g., all members of a species share the same general brain structure, which is partly shaped by performance-efficiency trade-offs). Sometimes, trade-offs can be addressed by deploying multiple subsystems or algorithms, each designed to privilege a different property. For example, evolution has endowed mammals with two distinct neural circuits to deal with predation threats: a rapid but inaccurate system that processes sensory information in the amygdala and triggers an immediate fear response, and a slow but accurate system that relies on cortical processing (Chittka et al., 2009). Similarly, reinforcement learning can take place through *model-free* or *model-based* algorithms. Model-free algorithms are computationally simpler and therefore more rapid, but also less accurate (and less flexible in response to changing conditions) than their model-based counterparts. The brain seems to make use of both types of algorithms, which may be instantiated in partially distinct neural regions (Daw & Dayan, 2014). Of course, the coexistence of multiple subsystems with different properties raises the higher-order problem of if and how to combine their outputs or "arbitrate" between them (e.g., Keramati et al., 2011).

On a shorter time scale, cognitive systems may dynamically adjust their functioning parameters depending on contextual factors, internal states, or the specific characteristics of a task. For example, raising the decision threshold of a decision-making process can increase accuracy at the expense of speed. (A similar result can be obtained by altering the balance between competing subsystems, e.g., by giving priority to slower and more accurate mechanisms.) Thus, foraging animals often respond to cues of predation risk by shifting to slower, more careful search strategies (Chittka et al., 2009). The optimal allocation of time between exploration and exploitation also depends on a wide array of factors, from the value and distribution of potential rewards to the predictability of the environment (Mehlhorn et al., 2015). Finally, alternative design solutions to functional trade-offs may be instantiated in stable patterns of individual differences: to illustrate, animals with "bold" personalities tend to make faster and less accurate decisions (Sih & Del Giudice, 2012), and individuals with higher intelligence and working memory capacity tend to devote more time to exploration, possibly because they benefit more from the additional information that they gain (Mehlhorn et al., 2015).

3.4. From Functional Trade-Offs to Empirical Correlations

All else being equal, functional trade-offs between competing properties of a system can be expected to give rise to negative correlations between indicators of those properties. For example, when participants are given stronger incentives to respond quickly in a decision task, the accuracy of their responses decreases accordingly. Varying the strength of the incentive should produce a negative correlation between response speed and accuracy, mirroring the underlying trade-off (Chittka et al., 2009; Heitz, 2014). However, all else may *not* be equal. For example, some participants in a group may process information at a much higher rate than others. Even when those participants respond quickly, their accuracy may still be higher than those of participants with a low processing capacity. It is also plausible that, under the same testing conditions, individuals with lower processing capacity will experience more severe trade-offs than those with higher capacity. If individual differences in processing capacity are large enough, they may significantly attenuate (or even reverse) the expected negative correlation between speed and accuracy.

An intriguing empirical finding in this respect is that people with higher general intelligence (as measured for example by IQ tests) tend to show *lower* neural activity and glucose consumption in the brain while solving cognitive tasks, especially if the tasks in question are relatively easy (Neubauer & Fink, 2009). This empirical pattern seemingly contradicts the trade-off between performance and energetic efficiency that would be expected based on the biophysics of information processing in neurons. However, the contradiction may be only apparent: as it turns out, the integrity of white matter fibers is associated with both higher intelligence and reduced neural activity, likely because intact and better-insulated fibers transmit electrical signals with higher fidelity (higher signal-to-noise ratio) and energetic efficiency (Penke et al., 2012; Warbrick et al., 2017). Confounding factors may operate between individuals—as in the case of white matter integrity—but also *within* the same individual over time: if state variables such as sleepiness or hormone levels affect performance, they can easily mask the existence of trade-offs at the group level (e.g., between an individual's average speed and its average accuracy; see Careau & Wilson, 2017).

The general point that empirical correlations may not mirror the underlying trade-offs is illustrated in Figure 3. Functional trade-offs may or may not give rise to negative correlations between empirical variables, because the existence of additional factors may easily lead to null or positive correlations that mask the underlying functional relationships. To reveal the underlying trade-offs, one needs to experimentally or statistically control for the effects of the confounding factors (Careau & Wilson, 2017). The distinction between a functional trade-off between two traits and a negative correlation between the same traits has been discussed most extensively in evolutionary biology. While organisms face multiple trade-offs in the allocation of resources to different life history traits (e.g., growth vs. production of offspring), individual variation in the availability of resources is often large enough to obscure the trade-offs at the phenotypic level (Reznick et al., 2000; Roff & Fairbairn, 2007; van Noordwijk & de Jong, 1986). This principle is obviously relevant to the analysis of trade-offs between multiple functional properties. For example, variation in resource availability (e.g., energy) may modulate existing trade-offs between performance and robustness: to the extent that performance and robustness

trade off against efficiency, it may be possible to increase both at the same time by relaxing the efficiency constraints on the system. We discuss this issue in more detail in a later section.

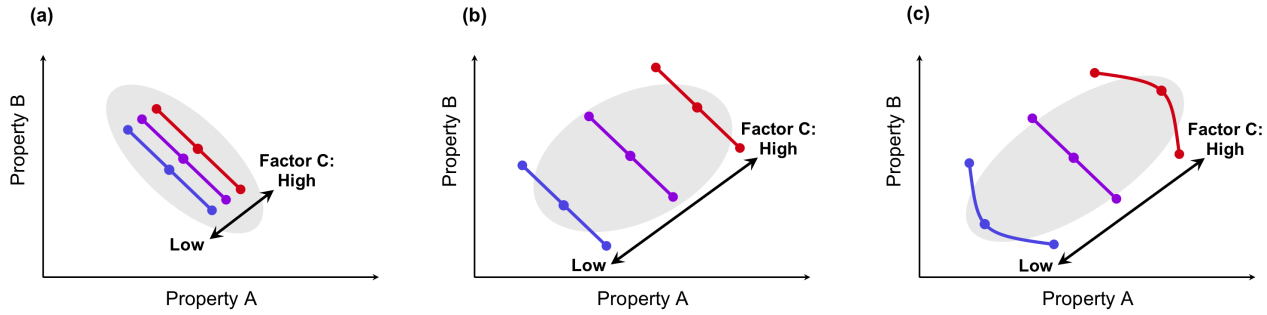

Figure 3. A functional trade-off between properties A and B (e.g., accuracy and speed) may give rise to different empirical correlations depending on variation in factor C (e.g., energetic resources). In panel (a), the overall correlation between A and B is negative, mirroring the underlying trade-off (negative slope of the lines). In panel (b), the effect of C is large enough that the overall correlation becomes positive, even if the same trade-off applies at any particular level of C. In panel (c), factor C also modulates the strength of the trade-off between A and B: the trade-off becomes stronger at lower levels of C (convex blue line) and weaker at higher levels of C (concave red line).

4. Robustness

Robustness is the ability of a system to maintain performance in the face of perturbations. There are many possible kinds of perturbations, both external (e.g., physical damage, extreme events that exceed the system's operating range) and internal (e.g., component failures, conflicts between subsystems). Cognitive systems are particularly exposed to perturbations caused by information corruption or *noise.* Noise can arise at any stage in the flow of information, including the system's input (e.g., sensors, neural connections), internal processing mechanisms, and output (e.g., inaccurate effectors), as well as in the environment in the form of stochastic fluctuations, sampling error, and unreliable cues to the true state of the world (Dayan, 2012; Flack et al., 2012; Gluck et al., 2012; Kitano, 2007).

Because noise and other perturbations are ubiquitous, the design of cognitive systems is heavily shaped by robustness demands. As we discuss in section 4.3, enhancing a system's robustness against one type of perturbation often generates new fragilities, which can then be addressed by additional mechanisms in a potentially never-ending cycle. The constant pressure for robustness contributes to the tendency for both natural and technological systems to become increasingly more complex over time (Anderson & Doyle 2010; Carlson & Doyle, 2002).

A system's robustness can be enhanced in a variety of ways (Anderson & Doyle 2010; Flack et al., 2012, Krakauer, 2006). A common design strategy to buffer the system against damage and component failure is to incorporate a degree of redundancy, with multiple units performing identical or overlapping tasks (e.g., processors in a multi-processor architecture; two brain hemispheres). The coexistence of alternative pathways or mechanisms that address the same task in complementary ways (e.g., fast vs. accurate processing) is a variation on this

principle. To further increase resistance to damage, such subsystems can be modularized (i.e., rendered functionally dissociable to a large degree) and/or compartmentalized in space. A related way to deal with errors and computational noise is to distribute a single information processing task among multiple units. Examples are natural and artificial neural networks, or statistical methods that make ensemble predictions by aggregating the results of a large number of semi-independent models (see Hastie et al., 2009). Distributed and ensemble processing can increase robustness by smoothing out noise and stochastic fluctuations, but also by buffering the system's functionality against the loss of individual units (e.g., neurons). All these strategies involve forms of *multiplicity,* whereby robustness is obtained through the coordinated activity of partially autonomous units (Dayan, 2012; Frank, 2008; Krakauer, 2006).

Multiplicity is not the only path to robustness. The quality of a system's components can be improved with better materials, increased precision, and more stringent quality control (e.g., selective elimination of defective neurons in the brain, or massive overproduction of neurons and synapses in early development followed by pruning). A basic engineering strategy is to build an extra safety margin in the components after the normal range of fluctuations has been calculated, so that the system will maintain functionality in the face of rare events that exceed the expected range. Biological organisms also encounter uncommon, extreme events that threaten the organism's integrity (e.g., heat shock, starvation, asphyxia). If such events are not exceedingly rare, biological systems can evolve "safety margins" around their normal range of operation as insurance against catastrophic failure. For tissues such as muscle, bone, and lung, safety factors have been estimated in the order of 2-10 times the normal range; neural circuits seem to possess a similar amount of excess capacity (Sterling & Freed, 2007; note that these estimates should be taken with a grain of salt, given the difficulty of assessing all the relevant variables). Increasing safety margins is a passive means to enhance robustness—but similar results can be achieved with the use of active processes. Errors in computation can be addressed by specialized error correction mechanisms; a system can be equipped with sensors that detect threats, perturbations, and malfunctions, making it possible to avoid them or counteract their effects.

An especially widespread strategy to counteract perturbations is to include feedback loops in the system (Anderson & Doyle 2010; Bechhoefer, 2005; Krakauer, 2006). Feedback controllers track the state of the system over time, correcting discrepancies between the desired and actual state as they arise. If the behavior of the system and the effect of perturbations can be modeled with some accuracy, feedback control can be supplemented with feedforward mechanisms that anticipate disturbances and correct them proactively instead of reactively (see Albertos & Mareels, 2010; Bechhoefer, 2005). Both feedback and feedforward control are extensively implemented in the architecture of the brain (e.g., Franklin & Wolpert, 2011; Wolpert et al., 2003; Yuste, 2015). Finally, the potentially catastrophic impact of rare outlier events ("black swans") can be attenuated by pessimistic decision-making strategies that are biased toward expectations of worst-outcome scenarios (Dayan, 2012; Gluck et al., 2012). Of course, all these robustness mechanisms have costs, and cannot be implemented without sacrificing other desirable properties of the system.

4.1. Trade-Offs Between Performance and Robustness

It is a common observation in systems biology and engineering that systems optimized for high performance on a certain task usually exhibit fragilities to perturbations, giving rise to *performance-robustness trade-offs* (Kitano, 2007, 2010; see Figure 2f). The evolution of modularity is a case in point. Individual modules become highly specialized for specific tasks or functions, so that performance increases (as well as efficiency); however, if a module is damaged the larger system may completely lose the ability to perform the corresponding task, with potentially dramatic consequences. Thus, selection for robustness may often work against the modularization of tasks, whereas selection for performance and efficiency tends to promote it (Rueffler et al., 2012; see also Clune et al., 2013).

An example from cognitive science is the fact that participants usually exhibit suboptimal performance in decision tasks (e.g., they fail to maximize the rate of rewards as predicted by optimality models). The participants' failure to achieve maximum performance can often be explained by their uncertainty about the parameters of the task (a form of noise), which prompts the adoption of robust worst-case strategies (Holmes & Cohen, 2014). More generally, pessimistic strategies that protect a system against the occurrence of rare outliers tend to diminish performance under normal operating conditions; they may also have significant opportunity costs, since risk-averse decision makers are more likely to forfeit favorable occasions when they unexpectedly present themselves (Dayan, 2012, Gluck et al., 2012). Of course, the performance costs associated with pessimistic strategies may be compensated for by the ability to avoid potentially disastrous mistakes. Design strategies that rely on multiplicity can also impair performance to some extent. Most notably, when multiple subsystems address the same task, the problem arises of how to arbitrate between their outputs. Mathematical models show that, since computational resources are always constrained, even a system designed to perform optimally may still experience systematic conflicts among its subsystems (Livnat & Pippenger, 2006). In less-than-optimal systems, internal conflicts between subsystems with divergent outputs have the potential to compromise performance, giving rise to inconsistent and possibly maladaptive results (Hagen et al., 2012).

Trade-offs between performance and robustness include the ubiquitous phenomenon known as *bias-variance trade-off* (Geman et al., 1992; Hastie et al., 2009). The classic formulation of the trade-off arises in frequentist statistical inference: while more complex models with many parameters and/or a complex functional form (e.g., a high-order polynomial regression model) provide a closer fit to the data and less biased estimates of the model parameters, they are more easily swayed by noise and sampling error (overfit), with the result that their estimates vary widely from one sample to the next (high variance). Simpler models (e.g., a linear regression model with few predictors) underfit the data and introduce systematic errors (high bias), but their estimates are more robust and less variable across samples, which considerably increases their usefulness for out-of-sample prediction. In other words, more complex models are less robust against noise (crucially including sampling error), but also less biased as inference tools and—as we discuss in section 5—more flexible in the range of data they can potentially fit. The scope of the bias-variance trade-off is not limited to statistical modeling: the same principles apply whenever a cognitive system attempts to learn or predict from data (Austerweil et al., 2015; Brighton & Gigerenzer, 2015; Gigerenzer & Brighton, 2009).

For instance, studies of concept formation show that exemplar-based concepts (in which all the individual examples of the concept are retained in memory) show low bias but high variance when they are used to classify a novel object, whereas prototype-like concepts (which are abstracted from individual examples through averaging) reduce classification variance at the cost of larger biases (Briscoe & Feldman, 2011).

From a certain perspective, absence of bias and low variance can be viewed as different components of accurate performance, roughly corresponding to trueness and precision (section 3.1). The relative importance of the two components depends on whether the goal of the system is limited to correct inference or involves future prediction. At the same time, biased inference can significantly increase a cognitive system's resistance to noise, and thus play a crucial role in enhancing its robustness (Austerweil et al., 2015). The dual functional role of high-bias, low-variance inference illustrates the broader point that the boundary between performance and robustness is not always clear-cut, and may shift depending on the exact goal of a system. For another example, consider the functional role of signal-to-noise ratio in neural transmission (see Sterling & Laughlin, 2015). A high signal-to-noise ratio can be viewed as an aspect of robustness (as it reflects the ability to reduce the impact of stochastic perturbations), but also as an aspect of performance (as it permits more accurate input-output mappings). Sometimes the ambiguity may be resolved by specifying the level of analysis to which each property applies; however, it is important to remember that conceptual categories such as "performance" and "robustness" are meant to be heuristically useful rather than absolutely precise.

Whereas robustness is defined as the ability to maintain performance against perturbation, the concept of *antifragility* (Taleb, 2012) refers to systems that improve their performance in response to perturbations, at least within a certain range. More precisely, Taleb (2012) defined antifragile systems as those that exhibit convex sensitivity to perturbations, so that stronger perturbations lead to disproportionally larger improvements. In a general sense, the ability to use perturbations to enhance performance is a pervasive feature of systems that learn from their failures and errors, including many cognitive systems. A less intuitive case is that of *stochastic resonance*, a fascinating phenomenon whereby adding a small amount of noise amplifies a weak signal instead of degrading it (see Hänggi, 2002). Stochastic resonance occurs in nonlinear systems that involve the crossing of a threshold, as in the action potential of neurons. Effects consistent with stochastic resonance have been demonstrated in various aspects of sensory processing in humans and other animals (Moss et al., 2004). While stochastic resonance does not meet Taleb's narrow definition of antifragility, it does represent a partial exception to performance-robustness trade-offs. The exception is only partial because it only applies to small amounts of noise: when noise exceeds the optimal intensity for the system, it begins to degrade performance in the usual way (Hänggi, 2002).

4.2. Trade-Offs Between Efficiency and Robustness

Most robustness strategies based on multiplicity require additional and potentially costly machinery—backup components, redundant subsystems, distributed units—as well as the energy to build, operate, and maintain it (Frank, 2008; Gluck et al., 2012). The reduced efficiency of complex and/or redundant systems may itself constitute a source of fragility with respect to sudden shortages of energy or other resources (Kitano, 2007). For example, the brain needs to

maintain a steady energetic supply, which in humans accounts for about 20% of the resting metabolic rate in adults and up to 65% in children (Kuzawa et al., 2014). Vulnerability to temporary food shortages during development seems to have been an important factor limiting the evolution of larger brains in mammals (Isler & van Schaik, 2009).

Depending on the architecture of a system, robust processes that rely on ensemble computations (e.g., running a sequence of models to average their outputs) may also take more time to produce a result (time inefficiency). Improving component quality, purging the system of defective units, and adding specialized mechanisms such as sensors and feedback control loops are similarly demanding in terms of resources. These pervasive trade-offs with efficiency have been labeled *robustness-resource trade-offs* (Kitano, 2007; see Figure 2c). An interesting example from the psychological literature is the distinction between two modes of cognitive control during goal-oriented tasks, labeled *proactive* and *reactive*. These two modes of control are thought to be implemented by distinct neural mechanisms, and their relative balance may contribute to stable individual differences in behavior (Botvinick & Braver, 2015; Braver, 2012; Braver et al., 2009; Coppens et al., 2010; Del Giudice, 2015; Huang et al., 2017). Proactive control depends on feedforward regulation: goal-related information is actively maintained in the attentional focus so as to anticipate or prevent interferences ("early selection"). In contrast, reactive control employs a form of feedback regulation: attention is recruited as needed after interferences are detected, so that corrective actions can be taken ("late correction"). While proactive control is more robust against disturbances, it is also more computationally demanding, and requires more attentional resources and working memory space. (As we discuss in section 5.3, proactive/feedforward control strategies also tend to be less flexible).

While robustness-resource trade-offs are very common, they are not always unavoidable; in some cases, a suboptimal system can be redesigned to use the same components in a more robust configuration (Khammash, 2016; Kitano, 2010). Also, it is often the case that design strategies that increase the overall robustness of a system (e.g., feedback loops, distributed computation) make it less sensitive to the quality or accuracy of its individual components. If so, the ability to use "sloppier" or lower quality components may contribute to reduce the system's cost (i.e., increase its resource efficiency; Anderson & Doyle 2010; Flack et al., 2012). This principle likely applies to the brain, where individual neurons transmit information with limited precision and introduce significant amounts of noise—not least owing to the energetic trade-offs discussed in a previous section. However, the structure of brain circuits can shape patterns of correlations among neurons in ways that substantially reduce the impact of noise on neural coding, thus enhancing the performance and robustness of the system as a whole (e.g., Zylberberg et al., 2016).

4.3. Trade-Offs Between Aspects of Robustness

Noise and perturbations come in many forms, and no system can be rendered immune to all possible disturbances. Many design strategies for robustness involve the introduction of additional components—sensors, feedback loops, backup subsystems, error correction processes—which, as noted earlier in this section, inevitably bring about new points of fragility. For instance, the same inhibitory circuits that stabilize brain activity through negative feedback can trigger epileptic seizures when they are perturbed in particular ways (Suffczynski et al.,

2004). The resulting vulnerability to uncontrolled oscillations is addressed by additional robustness mechanisms, including populations of astrocytes (a type of glial cells) that collectively regulate neural transmission and control the balance between excitation and inhibition in critical brain circuits. However, failure of this higher-order feedback system can severely destabilize the system; accordingly, astrocyte abnormalities play a significant role in the etiology of epilepsy (Amiri et al., 2011; Steinhäuser et al., 2016).

Observation of natural and artificial systems indicates that, as a general rule, the more a system depends on an intricate network of communication and control mechanisms, the more it is exposed to catastrophic failure if those mechanisms fail or are hijacked (Anderson & Doyle, 2010; Carlson & Doyle, 2002; Kitano, 2010). Paradoxically, systems that are especially well optimized to resist a specific kind of perturbation tend to become more vulnerable to unanticipated or rare events. The general principle that enhancing a system's robustness against one type of perturbation often generates fragilities to other types of perturbation is summarized by the phrase "robust yet fragile" (Figure 4); the resulting trade-offs have been labeled *robustness-fragility trade-offs* (Carlson & Doyle, 2002; Kitano, 2007). Robust-yet-fragile effects can easily determine spirals of increasing complexity over time, both in the design of technological systems and in the evolution of natural ones (Anderson & Doyle 2010). As organisms started using oxygen to extract energy from nutrients, the problem of avoiding damage from oxygen toxicity was addressed by the evolution of regulatory systems that maintain stable, narrowly controlled O_2 concentrations. Failure of these regulatory systems can be fatal to the organism—hence the evolution of multiple layers of physiological control with considerable redundancy. However, a stable internal environment becomes extremely favorable to parasites, prompting the evolution of complex immune systems with their own fragilities (including the problem of avoiding autoimmunity), and so on in a never-ending chain of new problems and solutions (the example is from Csete & Doyle, 2002).

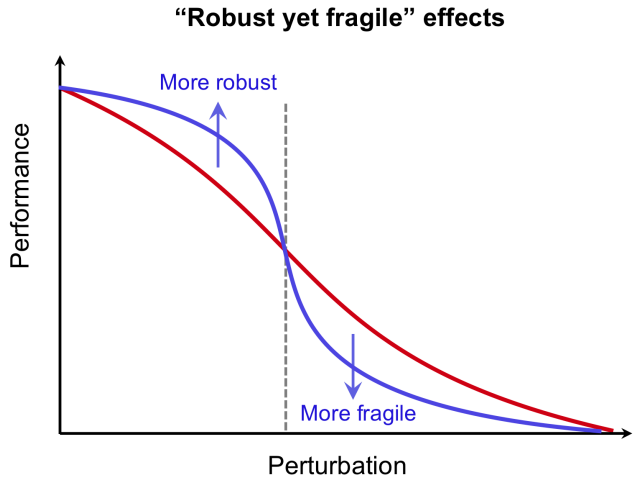

Figure 4. Schematic illustration of robust yet fragile effects. Enhancing a system's robustness to perturbations of a certain kind or within a certain range (left side of the figure) leads to increased fragility to perturbations of a different kind, or outside the range (right side of the figure).

While functional complexity can give rise to vulnerabilities, it would be a mistake to think that less complex systems are immune from robustness-fragility trade-offs. In fact, a classic case of robust-yet-fragile effect is the "conservation of fragility" in controllers based on negative feedback (Csete & Doyle, 2002; see Bechhoefer, 2005; Khammash, 2016). The goal of a feedback controller is to eliminate disturbances, which in this context mean deviations from the reference value (e.g., changes away from the set temperature in a thermostat). The system's performance can be regulated by adjusting the intensity of its response to deviations or *gain*. A higher gain improves the controller's ability to eliminate disturbances below a certain frequency; however, disturbances above that frequency are not reduced but *amplified*, and each increase in low-frequency stability (robustness) is exactly compensated by an equal increase in high-frequency instability (fragility). Consider the scenario of a thermostat connected to a heating/cooling unit, with the goal of keeping a room at a fixed set temperature. A thermostat that is extremely effective at canceling out slow temperature changes (e.g., between night and day) will break into uncontrolled oscillations if exposed to changes above a critical frequency (e.g., if another heater in the room is turned on and off every few minutes). In the case of feedback control, the conservation of fragility is an exact phenomenon that can be rigorously formalized; Kitano (2007) speculated that the same principle may apply more generally to biological systems that face a sufficiently broad range of perturbations, although this conjecture has yet to be thoroughly explored.

A subtler manifestation of robustness-fragility trade-offs is the fact that a system's robustness may be paradoxically enhanced by increasing the fragility of some of its components (Krakauer & Plotkin, 2004). This hierarchical version of the trade-off has many applications in engineering; a simple example is that of an electrical fuse that, precisely by virtue of its extreme vulnerability, protects the whole circuit from dangerous overcurrents. More generally, so-called "sacrificial parts" are engineered to fail first when they encounter perturbations that could compromise the functionality of the entire system. As well, when robustness is implemented through quality control (e.g., programmed cell death of defective neurons), fragility in the individual components can greatly contribute to the effectiveness of the selection process.

Robustness-fragility trade-offs are a major focus of interest in systems biology and engineering, but have been largely neglected in neuroscience and cognitive science. This gap is also a major opportunity for research: our understanding of neural and cognitive systems could be greatly improved by explicit consideration of how mechanisms that enhance robustness also give rise to specific fragilities, and—conversely—how apparent vulnerabilities may be the price to pay to guarantee robustness in other aspects of the system's performance. One area of research where these ideas could be fruitfully applied is *error management theory*, an evolutionary approach to decision-making that seeks to explain apparent cognitive biases (e.g., overconfidence in one's chances of success) as adaptive strategies that balance the risk of committing alternative types of errors (e.g., false positives vs. false negatives) when those errors have different fitness costs for the organism (Johnson et al., 2013). In the present framework, trade-offs between alternative types of decision errors can be framed as trade-offs between different aspects of performance. However, to the extent that decisions are influenced by noise, computational failures, and other perturbations (see Costello & Watts, 2014; Hilbert, 2012),

adaptive error management is also likely to entail trade-offs between different aspects of robustness/fragility.

5. Flexibility

Of the functional properties examined in this paper, flexibility is the hardest to pinpoint with precision; even in the scientific literature the term is often employed intuitively, without an explicit definition. Generally speaking, a system or organism is regarded as flexible if it can perform in a broad range of conditions and/or successfully adjust to changes and novelties in its operating environment (e.g., Coppens et al., 2010; Gluck et al., 2012; Liljenström, 2003). For cognitive systems, we suggest that flexibility can be recast as the ability to maintain performance (i.e., produce the intended input-output relationship) over a broad range of inputs, potentially including novel or unanticipated ones. We construe inputs broadly to include variation in operating conditions, domains of application, task demands, and so on.

Note that this working definition of flexibility focuses on the "signal" component of inputs, in contrast with robustness which is defined in relation to noise and perturbations. The definition highlights the main challenge faced by flexible systems—that is, distinguishing between novel/unusual but acceptable inputs (to which the system should adjust) and noise or other perturbations (that should be rejected or eliminated). While separating the signal from the noise is a universal cognitive problem, the task is easier for inflexible systems that accept a narrow range of inputs and discard much potentially valid information as noise. To illustrate, consider two hypothetical military systems designed to recognize different models of aircraft from pictures. The first can only recognize aircraft within a fixed set: each picture is matched to one of the models that are already known to the system (or a subset of the best matching ones). The second system is more flexible and has the capability of adding new models to the set as it encounters them. This system can potentially learn about the existence of previously unknown aircraft (e.g., secret or experimental models); however, it faces the additional problem of deciding whether unrecognized aircraft are merely familiar ones that have been distorted by noise (e.g., because they are viewed from an unusual angle), or are genuinely new models that should be added to the set. In many contexts, the distinction between novel/unusual inputs and noise is difficult or impossible to make *a priori*; this leads to pervasive trade-offs between flexibility and robustness, as we discuss in detail in section 5.3.

In contrast with efficiency or robustness, there has been little systematic investigation of design strategies that promote flexibility. A recurring theme in the literature is the ability to quickly update the system's operating parameters or stored information (e.g., Daw & Dayan, 2014; Liljenström, 2003). For this reason, flexibility can be enhanced by accepting new inputs without filtering them and processing them in real time. Conversely, flexibility is markedly reduced when the system is controlled by feedforward processes that ignore or discount new inputs. An extreme example of the latter is offered by defensive reflexes (e.g., retracting one's hand from a burning object), which once triggered tend to be carried out inflexibly and with little room for correction (Albertos & Mareels, 2010; Del Giudice, 2015). Less intuitively, slow-acting feedback processes can also stabilize a system, locking it in the current state and reducing the influence of new inputs. There are a number of neural feedback mechanisms (e.g., facilitation by recurrent excitatory synapses) that seem to play this role in the stabilization of memory traces

(Pereira & Wang, 2014). At a more abstract level of analysis, investing time in exploration (vs. exploitation; Mehlhorn et al., 2015) may contribute to increase the future flexibility of the system by broadening the range of inputs it is exposed to, as well as gathering information that can be stored and used later to improve performance in the face of change and novelty.

5.1. Trade-Offs Between Performance and Flexibility

The tension between performance and flexibility is captured by the adage "Jack of all trades, master of none" (Figure 2e). In behavioral ecology, this concept has been explored in the study of *generalist-specialist trade-offs.* Generalists are species or individuals who occupy a broad range of ecological niches (e.g., multiple food sources and foraging strategies), whereas specialists are restricted to a particular niche. A generalist lifestyle requires higher neural flexibility, including the ability to process different types of cues and keep track of more information about the environment. Because neural and cognitive resources (e.g., attention, memory) are inherently limited, specialists are expected to perform better than generalists when they operate in their particular niche—for example by making fewer errors and more accurate decisions during foraging (Bernays & Wcislo, 1994; Dall & Cuthill, 1997; Tosh et al., 2009). This prediction has been supported in a number of studies (mainly involving insects), although the evidence is not unequivocal (e.g., Bernays et al., 2004; Tapia et al., 2015; Tosh et al., 2011; Wee & Singer, 2007). Theoretical models suggest that the evolution of eco-cognitive specialization may require particular conditions, such as a low cost of decision errors (see Tosh et al., 2009, 2011).

In human psychology, the possibility of cognitive trade-offs on a generalist-specialist axis has not been widely investigated. Perhaps the closest analogue is "Spearman's law of diminishing returns," the empirical observation that cognitive abilities become more strongly differentiated at higher levels of general intelligence (Spearman, 1927; see Jensen, 2003). People with low intelligence tend to show uniformly poor performance across tasks involving different abilities (e.g., verbal, visual, spatial, reasoning); in contrast, people at the high end of the intelligence range are more likely to perform significantly better in some areas than others, suggesting a higher degree of cognitive specialization (Blum & Holling, 2017; Molenaar et al., 2017). More speculatively, levels of cognitive integration versus differentiation may correlate with personality traits, and partly reflect individual differences in evolved reproductive strategies (see Woodley, 2011). As with generalist-specialist trade-offs in ecology, the phenomenon of increasing differentiation at higher ability levels may depend on specific features and constraints of human cognition, and may not apply to other organisms—or even artificial agents (Hernández-Orallo, 2016). Finally, findings in neuroscience suggest that individuals who train for specialized performance may sacrifice some cognitive flexibility in the process. Most notably, Maguire et al. (2006) found that London taxi drivers (who need to learn complex spatial representations of the city) had an enlarged mid-posterior hippocampus compared with bus drivers (who follow constrained routes). However, they performed worse than bus drivers when they had to learn new spatial information, suggesting a loss of flexibility that correlated with diminished volume in the anterior hippocampus. In total, performance-flexibility trade-offs are both important and relatively understudied, making this an especially promising topic for future research.

5.2. Trade-Offs Between Efficiency and Flexibility

In addition to performing better than their more flexible counterparts, specialized cognitive systems that process a narrow range of inputs often show improved efficiency (Figure 2b). In ecology, there is evidence that specialists make faster decisions than generalists when faced with the same tasks (e.g., Bernays & Funk, 1999), and some theories of specialist advantages in cognition emphasize speed as well as accuracy (Bernays & Wcislo, 1994; Dall & Cuthill, 1997). To the extent that flexible cognition requires more intensive processing and a more spacious memory, one can also predict that—all else being equal—ecological specialists should have larger, more complex, and/or more energetically expensive brains than generalists. While the comparative study of brain structure is fraught with methodological difficulties (see Healy & Rowe, 2007), there is initial evidence that species that inhabit broader ecological niches tend to have larger and more complex brains, all else being equal (Lefebvre & Sol, 2008).

In the area of learning and decision-making, a key trade-off between efficiency and flexibility emerges in the contrast between model-based and model-free algorithms (Daw & Dayan, 2014). Model-based algorithms build an explicit representation of the state of the world and the likely consequences of each decision, instead of simply keeping track of the most successful response in each situation. The computational and memory requirements of model building can be formidable; as a result, model-based algorithms are slower and require considerably more cognitive resources. At the same time, the explicit models on which these algorithms depend are extremely easy to update with new information: it is sufficient to change a single critical value in the model (e.g., the probability of a certain state, or the expected consequence of a given choice) to immediately adjust the entire behavior of the system, without the need to laboriously re-learn what the best responses are in each condition. The superior efficiency of model-free algorithms comes at the price of rigid, inflexible behavior once learning has occurred (Daw & Dayan, 2014; Keramati et al., 2011). In a similar vein, learning mechanisms that favor exploration over exploitation (Dayan & Daw, 2008; Mehlhorn et al., 2015) make less efficient use of time, but may gain in flexibility and improve the system's ability to deal with unexpected inputs. Another classic finding in psychology is that practicing skills to the point of automaticity leads to dramatic increases in efficiency (speed, low use of attentional resources; see Logan, 1985). At the same time, highly automatized behaviors (e.g., driving a car) become more difficult to adjust when conditions change abruptly (e.g., when switching from right-hand to left-hand traffic).

A somewhat different illustration of the same basic trade-off is provided by *fast-and-frugal heuristics*, a class of simple algorithms that make rapid decisions (fast) by discarding most of the available information and employing only a few cues from the environment (frugal). Fast-and-frugal heuristics are defined by their efficiency; in many real-world conditions, they can outperform more complex algorithms that make full use of the available information, such as linear regression (see Neth & Gigerenzer, 2015; Todd et al., 2012). The key to the success of simple heuristics is their "ecological rationality:" each particular heuristic is matched to a specific kind of environment, and works by exploiting ecological regularities while avoiding overfitting (i.e., minimizing variance) by virtue of its computational simplicity. For example, heuristics that base decisions on a single cue while ignoring all other information—for example, only looking at meal price to choose the best restaurant—perform well when cues are correlated

and thus partly redundant, but some cues have much higher validity than others (Gigerenzer & Brighton, 2009). In other words, the efficiency of fast-and-frugal heuristics (as well as their robustness; see section 5.3) is partly a function of their lack of flexibility. Since each heuristic is tuned to specific characteristics of the environment and/or task, the main challenge for the decision-maker becomes that of selecting the appropriate heuristic for a given situation (Lieder & Griffiths, 2017; Marewski & Schooler, 2011). Interestingly, recent work indicates that fast-and-frugal heuristics can themselves be outperformed by Bayesian models that use all the available information while heavily discounting some of it in order to match the structure of the environment or task, rather than discarding it altogether as heuristics do (Parpart et al., 2017). These models are somewhat more flexible than fast-and-frugal heuristics, but markedly less efficient owing to their computational complexity.

5.3. Trade-Offs Between Robustness and Flexibility

As discussed in section 4, a common strategy to increase a system's resistance to noise and perturbations is to make it less sensitive to fluctuations in the input, which are rejected or eliminated. Since reducing noise effectively narrows the range of acceptable inputs, it is often the case that enhancing a system's robustness simultaneously reduces its flexibility to some extent (Figure 2d). In psychology and neuroscience, this is often referred to as the *stability-flexibility dilemma,* with "stability" used as synonym for robustness as defined here (Goshke, 2000; Liljenström, 2003). Computational models of neural networks show that introducing a certain amount of spontaneously generated noise (e.g., through chaotic oscillatory dynamics) increases the network's responsivity to new inputs and prevents the system from getting stuck in non-optimal states—all while reducing its stability (Liljenström, 2003). In a neural model of short-term memory, slowing down the time course of feedback mechanisms such as recurrent excitatory facilitation increases the stability of memory traces, making them more resistant to noise and interference. However, the network also becomes harder to reset and update with new information, and hence less flexible (Pereira & Wang, 2014). Some authors have argued that a key function of dopamine in the brain is to regulate the trade-off between cognitive flexibility—which comes at the cost of distractibility—and robustness to interference, whose downsides are behavioral rigidity and lack of responsiveness to new information (Cools & D'Esposito, 2011; Hills, 2006). Specifically, dopaminergic activity in the striatum seems to promote flexibility, whereas prefrontal dopamine increases stability and persistence (Boot et al., 2017).

Similar considerations apply to the distinction between proactive and reactive control mechanisms that was introduced in a previous section. Proactive control is more robust against perturbations and interference, but also less flexible; its anticipatory, feedforward nature makes it hard to adjust behavior if environmental conditions change unexpectedly. Reactive mechanisms can easily respond to change, but are also more vulnerable to the effects of environmental noise (Braver, 2012; see also Tops et al., 2010). The balance between proactive control (robust but inflexible) and reactive control (flexible but fragile) may contribute to broader individual differences in behavior, which in the animal literature are captured by the distinction between proactive and reactive *coping styles* (defined as stable patterns of behavioral and physiological responses to challenges; see Coppens et al., 2010; Del Giudice, 2015). Note that, despite their conceptual overlap, the modes of control described in the cognitive literature and the coping styles studied in behavioral biology are not identical. For example, reactive control mechanisms

are regarded as more computationally efficient (Braver, 2012), whereas animals with reactive coping styles tend to engage in more accurate but less efficient exploration—they are slower, more thorough, and store more detailed information in memory (Sih & Del Giudice, 2012). Despite these differences, the two constructs are linked by a shared emphasis on the central trade-off between robustness and flexibility.

Fast-and-frugal heuristics manage to outperform more complex decision algorithms by minimizing variance—that is, maximizing robustness—at the cost of increased bias. In turn, the negative effects of bias are reduced by sacrificing flexibility and matching each heuristic to a particular kind of environment or task. This strategy succeeds when the input is noisy and uncertainty is high—for example when inference is carried out on small amounts of data, when the problem is exceedingly complex, and/or when the available information is unreliable. All these conditions magnify the benefits of robustness. When noise and uncertainty are low, more complex and flexible algorithms tend to outperform simpler heuristics (Gigerenzer & Brighton, 2009). In the next section, we discuss the effectiveness of fast-and-frugal heuristics in more detail as an illustration of simultaneous trade-offs between multiple functional properties.

6. Multiple Trade-Offs

Up to this point, we have focused our analysis on trade-offs between pairs of functional properties—performance versus efficiency, robustness versus flexibility, and so on. However, several of the examples we discussed involve multiple trade-offs that jointly define the system's design constraints. Ecological specialists are both faster and more accurate than generalists (Bernays & Wcislo, 1994; Dall & Cuthill, 1997); thus, the generalist-specialist trade-off entails a simultaneous trade-off of flexibility versus performance *and* efficiency. In the distinction between proactive and reactive control mechanisms, the trade-off is between the robustness of proactive control on the one hand, and the efficiency *and* flexibility of reactive control on the other hand (Braver, 2012; Mazza et al., 2018). Learning strategies that favor exploration sacrifice efficiency to increase both performance and flexibility (Mehlhorn et al., 2015).

In many cases involving multiple trade-offs, the original trade-off between two properties A and B can be partially overcome by changing a third property C, as illustrated in Figure 3. For example, trade-offs between performance and flexibility (such as those that underlie the generalist-specialist distinction) should become less stringent when time is abundant and organisms can engage in extensive exploration without significant costs. Or, increased energy availability (for example through better diet) may reduce efficiency constraints on brain structure and function, making it possible to simultaneously increase both the speed and the accuracy of neural transmission. In a previous section, we noted that IQ correlates with the integrity of white matter fibers; the resulting increase in overall efficiency likely contributes to explain why higher IQ predicts better cognitive performance but lower energetic consumption in the brain (Neubauer & Fink, 2009). Note that the effect of C on the trade-off between A and B can take two distinct forms. First, changing the level of C may increase both A and B simultaneously, but without altering the shape of the original trade-off (Figure 3b). Second, C may alter the underlying functional relation between A and B, so that the shape of the trade-off changes accordingly, for example becoming less severe (Figure 3c). In most of the examples discussed in this paper, the

evidence is insufficient to discriminate between the two scenarios; however, the distinction is conceptually important and could usefully inform future research in this area.

An especially intriguing illustration of a multiple trade-off was discussed by Vakulenko and Radulescu (2012) in their formal analysis of genetic networks, which may also have implications for the organization of neural networks. These authors showed that centralized network architectures can be both highly flexible (i.e., capable of switching between many different states) and highly robust to random perturbations; however, such networks must necessarily function at a slow rate. In this case, the trade-off between robustness and flexibility can be overcome by sacrificing time efficiency. For a different example, consider statistical methods that make ensemble predictions by fitting a large number of semi-independent models and averaging their results (Hastie et al., 2009). Each individual model has low bias but high variance (overfitting); when their results are averaged, the overfitting features of individual models cancel out, and the resulting prediction can have both extremely low variance *and* low levels of bias. What ensemble methods do is sacrificing efficiency (in terms of time, computation, and memory space) to overcome the trade-offs between performance and robustness that constrain the potential of individual models.

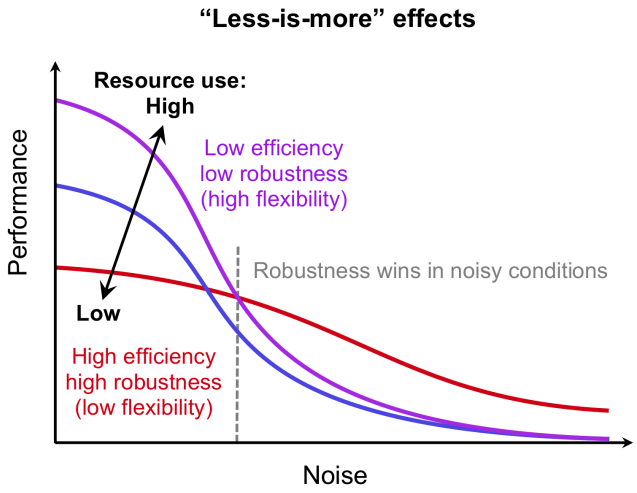

Figure 5. Less-is-more effects as triple trade-offs between efficiency, robustness, and flexibility. When conditions are sufficiently noisy, robust heuristics that use limited information and computational resources systematically outperform more complex and resource- or knowledge-intensive algorithms (right side of the figure). However, such efficient and robust heuristics lack flexibility, and perform well only when matched to a specific kind of environment and/or task.

Finally, looking at fast-and-frugal heuristics from the standpoint of functional trade-offs helps clarify a crucial but counterintuitive phenomenon. Under the right conditions, simple heuristics can be more accurate than more complex algorithms that use all the available information, even with unlimited time and computational resources; so that increasing the amount of resources devoted to computation actually results in *worse* performance. This violation of the performance-efficiency trade-off (or *effort-accuracy trade-off* in the authors' terminology) is an instance of what have been labeled "less-is-more effects" (Gigerenzer &

Brighton, 2009; Katsikopoulos, 2010). The violation is only apparent, however. The key to the superiority of fast-and-frugal heuristics lies in the combination of robustness—which boosts their performance in noisy conditions—and specialization for a particular kind of environment or task (ecological rationality). In other words, the lack of flexibility of these heuristics ultimately allows them to perform well despite their extreme efficiency. Figure 5 illustrates how less-is-more effects can be understood as stemming from a triple trade-off between robustness, efficiency, and flexibility (Figure 5).

7. Conclusions

To understand the design of cognitive systems, it is imperative to think clearly about trade-offs and their implications. In this paper, we have sought to organize a large and diverse literature on trade-offs, using four basic properties of functional systems, performance, robustness, efficiency, and stability as a frame of reference and organization (Figure 1). Drawing on the shared vocabularies of systems biology and engineering, the framework, we have presented abstract descriptions of several crucial design problems (Figure 2). Such descriptions can be used to make sense of specific trade-offs encountered in cognitive research, as well as to extract general principles and insights that can be usefully transferred across disciplines and topics, fostering cross-fertilization and suggesting new directions for investigation. As it should be clear from the paper, we are not suggesting *replacement* of models for specific trade-offs with more general descriptions; rather, we believe it can be useful to examine specific trade-offs in the context of the broader design problems they represent, so that causes, effects, connections, and commonalities may more easily emerge.

While the functional properties we examined in this paper are highly general, the scope of the framework is limited by its focus on single goal-oriented systems that perform an identifiable, relatively well-defined function. The framework does not directly apply to cases in which trade-offs occur between two or more cognitive systems with competing functions and no shared performance criterion. For example, trade-offs have been posited between “mechanistic” processes that model the behavior of predictable physical objects and “mentalistic” processes specialized to predict the behavior of intentional agents (Crespi & Badcock, 2008; see Baron-Cohen, 2009). By this hypothesis, enhanced abilities on one of the two process domains (such as increased visual-spatial skills in autism) often involve reduced abilities in the other, due in part to neurologically-based tradeoffs in the engagement of anticorrelated brain regions and networks (e. g., Jack et al., 2013; Crespi & Go, 2015). The question of how different mechanisms with competing and partially conflicting functions compete for the control of behavior and shared cognitive resources (e.g., working memory) cannot be addressed in the present framework, and will require further theoretical work before it can receive a satisfactory answer. Interactions among multiple cognitive systems are especially relevant to approaches that view the mind as a rich collection of efficient domain- or task-specific mechanisms, such as the “adaptive toolbox” model of heuristics in the field of ecological rationality (Todd et al., 2012) or the massive modularity hypothesis in evolutionary psychology (Barrett & Kurzban, 2006; Carruthers, 2006).

Another potentially fruitful direction for extending the framework is to consider collective cognition and the trade-offs it entails (Couzin, 2009). In collective cognition, information processing is distributed among multiple individuals who may have different

abilities and knowledge, but also partially different goals and biological interests (e.g., Conradt & Roper, 2009; Kameda et al., 2011). Typical problems that involve group decision-making include migration, foraging, predator avoidance, and nest building (Couzin, 2009; King & Sueur, 2011). In humans, cooperation and collective information sharing occur on a massive scale, prompting research on the effects of different learning strategies and social network structures (e.g., Barkoczi & Galesic, 2016). Research in this area has explored trade-offs between speed and accuracy in group decision-making (Couzin, 2009; Franks et al., 2003); between exploration and exploitation in the spread of innovations throughout social networks (Mason et al., 2008; Wisdom et al., 2013); and between performance in simple versus complex tasks in human groups that follow different learning rules (Barkoczi & Galesic, 2016). Intriguingly, some theoretical models suggest the existence of collective less-is-more effects, so that—under certain conditions—performance can be higher in groups whose individual members are less knowledgeable and/or competent (Luan et al., 2013).

The primary proximate sources of cognitive trade-offs, and variation among individuals and species in their forms and expression, are neural, neuroendocrine and bioenergetic systems, and the constraints that structure neurodevelopment. A key challenge in studies of the causes of cognitive trade-offs is thus to determine the genetic, neurological and developmental-level mechanisms of trade-offs in cognitive phenotypes (e. g., Heitz & Schall 2012). In human psychology, cognitive trade-offs and their mechanisms should contribute to explaining individual differences in aspects of personality, inelligence, memory, and other core phenotypes, given that individuals are expected to vary in traits that affect their positions along the dimensions and sides of tradeoffs (e.g., Crespi, 2015; Del Giudice, 2015; Sih & Del Giudice, 2012; Mehlhorn et al. 2015; Woodley, 2011). Psychopathology is another important area in which explicit trade-off thinking can be expected to yield useful insights. Extreme expression or dysregulation of life history and cognitive trade-offs have been implicated in the origin of several mental disorders (e.g., Cools et al., 2011; Crespi, 2015; Crespi & Badcock, 2008; Crespi & Go, 2015; Del Giudice, 2014). Some of those specific trade-offs may be fruitfully framed in the context of the broad functional properties in Figure 1; for example, pathologically high flexibility at the expense of robustness may characterize the risk for conditions such as psychotic disorders (schizophrenia, bipolar with mania) and attention-deficit/hyperactivity (Hills, 2006). Studies of psychopathology might benefit by focusing more on variation in trade-offs between affected individual and controls, compared with the characterization of trait deficits *per se*.

The analysis of cognitive trade-offs also draws attention to the issue of similarities and differences between biological and artificial intelligence systems with regard to trade-off architectures (e.g., Hernández-Orallo, 2016). For example, advanced neural networks optimized for accuracy and/or speed in object recognition can be surprisingly fragile against extremely small perturbations—sometimes consisting of a single pixel—that have been specifically engineered to "fool" their algorithms, but may be invisible to human observers (Akhtar & Mian, 2018; Moosavi-Dezfooli et al., 2017). This finding is consistent with the idea that demands for robustness—including against attacks and interference by other organisms—are a critical source of constraints on the performance of natural intelligence systems, and may loom larger on artificial systems as they are employed more often in critical real-world applications and subjected to malicious attacks (Akhtar & Mian, 2018). Other important constraints on performance likely stem from biological organisms' need to perform flexibly across a wide range

of tasks and novel inputs, with relatively little time for learning (see Edelman, 2016). Finally, future research on cognitive trade-offs in humans should inform the debate on cognitive enhancement by pharmacological and/or genetic means (Fox et al., 2017; Hills & Hertwig, 2011; Shulman & Bostrom, 2014). This is a complex issue whose myriad empirical, ethical, and policy ramifications cannot be meaningfully addressed without a deep understanding of the existing constraints on human cognition, their evolutionary underpinnings, and the costs and benefits that may result from novel interventions. In conclusion, we believe that the conceptual framework we have presented here will help to organize knowledge across wide-ranging, disparate areas of research, and we hope that it will find applications in a variety of domains, from cognitive science to psychiatry.

Acknowledgments

We are grateful to Peter Todd and two anonymous reviewers for their many insightful comments on an earlier draft of this paper.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30-60.

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv*, 1801.00553.

Albertos, P., & Mareels, I. (2010). *Feedback and control for everyone*. Heidelberg, DE: Springer.

Alderson, D. L., & Doyle, J. C. (2010). Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Transactions on Systems, Man, and Cybernetics A: Systems and Humans, 40,* 839-852.

Amiri, M., Bahrami, F., & Janahmadi, M. (2011). Functional modeling of astrocytes in epilepsy: A feedback system perspective. *Neural Computing and Applications, 20,* 1131-1139.

Attwell, D., & Laughlin, S.B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism, 21,* 1133–1145.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 187-208). New York, NY: Oxford University Press.

Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Nature Communications, 7,* 13109.

Baron-Cohen, S. (2009) Autism: the empathizing–systemizing (E-S) theory. *Annals of the New York Academy of Sciences*, *1156*, 68-80.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113,* 628-647.

Bechhoefer, J. (2005). Feedback for physicists: A tutorial essay on control. *Reviews of Modern Physics, 77,* 783-836.

Beer, R. D., & Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive Science, 39*, 1-38.

Bernays, E. A., & Funk, D. J. (1999). Specialists make faster decisions than generalists: Experiments with aphids. *Proceedings of the Royal Society of London B, 266,* 151-156.

Bernays, E. A., Singer, M. S., & Rodrigues, D. (2004). Foraging in nature: Foraging efficiency and attentiveness in caterpillars with different diet breadths. *Ecological Entomology, 29,* 389-397.

Bernays, E. A., & Wcislo, W. T. (1994). Sensory capabilities, information processing, and resource specialization. *Quarterly Review of Biology, 69,* 187-204.

Blum, D., & Holling, H. (2017). Spearman's law of diminishing returns. A meta-analysis. *Intelligence, 65,* 60-66.

Boot, N., Baas, M., van Gaal, S., Cools, R., & De Dreu, C. K. (2017). Creative Cognition and Dopaminergic Modulation of Fronto-striatal Networks: Integrative Review and Research Agenda. *Neuroscience & Biobehavioral Reviews, 78,* 13-23.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology, 66,* 83-113.

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

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences USA, 106,* 7351-7356.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research, 68,* 1772-1784.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance trade-off. *Cognition, 118,* 2-16.

Careau, V., & Wilson, R. S. (2017). Of Uberfleas and Krakens: detecting trade-offs using mixed models. *Integrative and Comparative Biology, 57,* 362-371.

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

Carruthers, P. (2006). The *architecture of the mind: Massive modularity and the flexibility of thought.* New York: Oxford University Press.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9,* 129-136.

Chittka, L., Skorupski, P., & Raine, N. E. (2009). Speed–accuracy trade-offs in animal decision making. *Trends in Ecology and Evolution, 24,* 400-407.

Clune, J., Mouret, J. B., & Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society of London B, 280,* 20122863.

Conradt, L., & Roper, T. J. (2009). Conflicts of interest and the evolution of decision sharing. *Philosophical Transactions of the Royal Society B, 364,* 807-819.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30,* 163-183.

Cools, R., & D'Esposito, M. (2011). Inverted-U–shaped dopamine actions on human working memory and cognitive control. *Biological Psychiatry, 69,* e113-e125.

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology, 36,* 98-113.

Coppens, C. M., de Boer, S. F., & Koolhaas, J. M. (2010). Coping styles and behavioural flexibility: Towards underlying mechanisms. *Philosophical Transactions of the Royal Society B, 365,* 4021–4028.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review, 121,* 463-480.

Couzin, I. D. (2009). Collective cognition in animal groups. *Trends in Cognitive Sciences, 13,* 36-43.

Crespi, B. (2015) Cognitive trade-offs and the costs of resilience. *Behavioral and Brain Sciences*, 38, 27-28.

Crespi, B., & Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain. *Behavioral and Brain Sciences*, *31*, 241-261.

Crespi, B. & Go, M. (2015) Diametrical diseases reflect evolutionary-genetic trade-offs: Evidence from psychiatry, neurology, rheumatology, oncology and immunology. *Evolution, Medicine, and Public Health*, *2015*, 216-253.

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science, 295,* 1664-1669.

Dall, S. R. X., & Cuthill, I. C. (1997). The information costs of generalism. *Oikos, 80,* 197-202.

Dayan, P. (2102). Robust neural decision making. In P. Hammerstein, & J. R. Stevens (Eds.), *Evolution and the mechanisms of decision making* (pp. 151-168). Cambridge, MA: MIT Press.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, and Behavioral Neuroscience, 8,* 429-453.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B, 369,* 20130478.

Del Giudice, M. (2014). An evolutionary life history framework for psychopathology. *Psychological Inquiry, 25,* 261-300.

Del Giudice, M. (2015). Self-regulation in an evolutionary perspective. In G. H. E. Gendolla, M. Tops & S. Koole (Eds.), *Handbook of biobehavioral approaches to self-regulation* (pp. 25-42). New York: Springer.

Dennett, D. (2009). Darwin's "strange inversion of reasoning". *Proceedings of the National Academy of Sciences USA, 106,* 10061-10065.

Doyle, J. C., & Csete, M. (2011). Architecture, constraints, and behavior. *Proceedings of the National Academy of Sciences USA, 108,* 15624-15630.

Edelman, S. (2016). The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence, 28,* 751-776.

Flack, J. C., Hammerstein, P., & Krakauer, D. C. (2012). Robustness in biological and social systems. In P. Hammerstein, & J. R. Stevens (Eds.), *Evolution and the mechanisms of decision making* (pp. 129-150). Cambridge, MA: MIT Press.

Fox, K. C., Fitz, N. S., & Reiner, P. B. (2017). The multiplicity of memory enhancement: Practical and ethical implications of the diverse neural substrates underlying human memory systems. *Neuroethics, 10,* 375-388.

Frank, S. A. (2008). Evolutionary dynamics of redundant regulatory control. *Journal of Theoretical Biology, 255,* 64-68.

Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron, 72,* 425-442.

Franks, N. R., Dornhaus, A., Fitzsimmons, J. P., & Stevens, M. (2003). Speed versus accuracy in collective decision making. *Proceedings of the Royal Society of London B, 270,* 2457-2463.

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology, 56,* 1–104.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4,* 1–58.

Gigerenzer, G., & Brighton, H. (2009). *Homo heuristicus*: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107-143.

Gluck, K. A., McNamara, J. M., Brighton, H., Dayan, P. Kareev, Y., Krause, J., … & Wismatt, W. C. (2012). Robustness in a variable environment. In P. Hammerstein & J. R. Stevens (Eds.), *Evolution and the mechanisms of decision making* (pp. 195-214). Cambridge, MA: MIT Press.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task-set switching. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 331- 355). Cambridge, MA: MIT Press.

Hagen, E. H., Chater, N., Gallistel, C. R., Houston, A., Kacelnik, A., Kalenscher, T., … & Stephens, D. W. (2012). Decision making: What can evolution do for us? In P. Hammerstein, & J. R. Stevens (Eds.), *Evolution and the mechanisms of decision making* (pp. 97-126). Cambridge, MA: MIT Press.

Hänggi, P. (2002). Stochastic resonance in biology: How noise can enhance detection of weak signals and help improve biological information processing. *ChemPhysChem, 3,* 285-290.

Harris, J. J., Jolivet, R., Engl, E., & Attwell, D. (2015). Energy-efficient information transfer by visual pathway synapses. *Current Biology, 25,* 3151-3160.

Hasenstaub, A., Otte, S., Callaway, E., & Sejnowski, T. J. (2010). Metabolic cost as a unifying principle governing neuronal biophysics. *Proceedings of the National Academy of Sciences USA, 107,* 12329-12334.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Healy, S. D., & Rowe, C. (2007). A critique of comparative studies of brain size. *Proceedings of the Royal Society of London B, 274*, 453-464.

Heitz, R. P. (2014). The speed-accuracy trade-off: history, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8,* 150.

Heitz, R. P. & Schall, J. D. (2012) Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*, 616-628.

Hernández-Orallo, J. (2016). Is Spearman's Law of Diminishing Returns (SLODR) meaningful for artificial agents? *Proceedings of the 22nd European Conference on Artificial Intelligence* (pp. 471-479). Amsterdam, NL: IOS Press.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138,* 211-237.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science, 30,* 3-41.

Hills, T. T., & Hertwig, R. (2011). Why aren't we smarter already: Evolutionary trade-offs and cognitive enhancements. *Current Directions in Psychological Science, 20,* 373-377.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences, 19,* 46 –54.

Holmes, P., & Cohen, J. D. (2014). Optimality and some of its discontents: Successes and shortcomings of existing models for binary decisions. *Topics in Cognitive Science, 6,* 258-278.

Huang, S., Zhu, Z., Zhang, W., Chen, Y., & Zhen, S. (2017). Trait impulsivity components correlate differently with proactive and reactive control. *PLoS ONE, 12,* e0176102.

Isler, K., & van Schaik, C. P. (2009). The expensive brain: A framework for explaining evolutionary changes in brain size. *Journal of Human Evolution, 57,* 392-400.

Jack, A.I., Dawson, A.J., Begany, K.L., Leckie, R.L., Barry, K.P., Ciccia, A.H. & Snyder, A.Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385-401.

Jensen, A. R. (2003). Regularities in Spearman's law of diminishing returns. *Intelligence, 31,* 95-105.

Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews, 46,* 30-43.

Johnson, D. D., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology and Evolution, 28,* 474-481.

Joint Committee for Guides in Metrology (2008). *International vocabulary of metrology - Basic and general concepts and associated terms (VIM).* http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf

Kameda, T., Tsukasaki, T., Hastie, R., & Berg, N. (2011). Democracy under uncertainty: The wisdom of crowds and the free-rider problem in group decision making. *Psychological Review, 118,* 76-96.

Katsikopoulos, K. V. (2010). The less-is-more effect: Predictions and tests. *Judgment and Decision Making, 5,* 244-257.

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

Khammash, M. (2016). An engineering viewpoint on biological robustness. *BMC Biology, 14,* 22.

King, A. J., & Sueur, C. (2011). Where next? Group coordination and collective decision making by primates. *International Journal of Primatology, 32,* 1245-1267.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics, 5,* 826-837.

Kitano, H. (2007). Towards a theory of biological robustness. *Molecular Systems Biology, 3,* 137.

Kitano, H. (2010). Violations of robustness trade-offs. *Molecular Systems Biology, 6,* 384.

Krakauer, D. C. (2006). Robustness in Biological Systems: A provisional taxonomy. In S. Deisboeck & J. Yasha Kresh (Eds.), *Complex systems science in biomedicine* (pp. 183-205). New York, NY: Springer.

Krakauer, D. C., & Plotkin, J. (2004). Principles and parameters of molecular robustness. In E. Jen (Ed.), *Robust design: A repertoire for biology, ecology, and engineering* (pp. 71-103). New York, NY: Oxford University Press.

Kurup, U., & Lebiere, C. (2012). What can cognitive architectures do for robotics? *Biologically Inspired Cognitive Architectures, 2,* 88-99.

Kuzawa, C. W., Chugani, H. T., Grossman, L. I., Lipovich, L., Muzik, O., Hof, P. R., ... & Lange, N. (2014). Metabolic costs and evolutionary implications of human brain development. *Proceedings of the National Academy of Sciences USA, 111,* 13010-13015.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research, 10,* 141-160.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology, 11,* 475-480.

Lefebvre, L., & Sol, D. (2008). Brains, lifestyles and cognition: Are there general trends? *Brain, Behavior and Evolution, 72,* 135-144.

Lennie, P. (2003). The cost of cortical computation. *Current biology*, *13*, 493-497.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science, 6,* 279-311.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124,* 762-794.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 2690-2798). Cambridge, MA: MIT Press.

Liljenström, H. (2003). Neural stability and flexibility: A computational approach. *Neuropsychopharmacology, 28,* S64.

Livnat, A., & Pippenger, N. (2006). An optimal brain can be composed of conflicting agents. *Proceedings of the National Academy of Sciences USA, 103,* 3198-3202.

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

Luan, S., Katsikopoulos, K. V., & Reimer, T. (2012). The "less-is-more" effect in group decision making. In R. Hertwig, U. Hoffrage, & the ABC Research Group (Eds.). *Simple heuristics in a social world* (pp. 319-342). New York: Oxford University Press.

Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus, 16,* 1091-1101.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118,* 393-437.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General, 137,* 422-433.

Mazza, V., Eccard, J. A., Zaccaroni, M., Jacob, J., & Dammhahn, M. (2018). The fast and the flexible: Cognitive style drives individual variation in cognition in a small mammal. *Animal Behaviour, 137,* 119-132.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... & Gonzalez, C. (2015). Unpacking the exploration–exploitation trade-off: A synthesis of human and animal literatures. *Decision, 2*, 191-215.

Molenaar, D., Kő, N., Rózsa, S., & Mészáros, A. (2017). Differentiation of cognitive abilities in the WAIS-IV at the item level. *Intelligence, 65,* 48-59.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal Adversarial Perturbations. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,* 86-94.

Moss, F., Ward, L. M., & Sannita, W. G. (2004). Stochastic resonance and sensory information processing: a tutorial and review of application. *Clinical Neurophysiology, 115,* 267-281.

Neth, H., & Gigerenzer, G. (2015). Heuristics: Tools for an uncertain world. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–18). New York, NY: Wiley.

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience & Biobehavioral Reviews, 33,* 1004-1023.

Parpart, P., Jones, M., & Love, B. (2017). Heuristics as Bayesian inference under extreme priors. *PsyArXiv*, doi:10.17605/OSF.IO/QKBT5

Penke, L., Maniega, S. M., Bastin, M. E., Hernández, M. V., Murray, C., Royle, N. A., ... & Deary, I. J. (2012). Brain white matter tract integrity as a neural foundation for general intelligence. *Molecular Psychiatry, 17,* 1026-1030.

Pereira, J., & Wang, X. J. (2014). A trade-off between accuracy and flexibility in a working memory circuit endowed with slow feedback mechanisms. *Cerebral Cortex, 25,* 3586-3601.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37,* 1-38.

Quax, R., Har-Shemesh, O., Thurner, S., & Sloot, P. (2016). Stripping syntax from complexity: An information-theoretical perspective on complex systems. *arXiv*, 1603.03552.

Reznick, D., Nunney, L., & Tessier, A. (2000). Big houses, big cars, superfleas and the costs of reproduction. *Trends in Ecology and Evolution, 15,* 421-425.

Roff, D. A., & Fairbairn, D. J. (2007). The evolution of trade-offs: where are we? *Journal of Evolutionary Biology, 20,* 433-447.

Rosenbloom, P. S., Laird, J. E., Newell, A., & McCarl, R. (1991). A preliminary analysis of the Soar architecture as a basis for general intelligence. *Artificial Intelligence, 47,* 289-325.

Rueffler, C., Hermisson, J., & Wagner, G. P. (2012). Evolution of functional specialization and division of labor. *Proceedings of the National Academy of Sciences USA, 109*, E326-E335.

Savage, J. E. (2008). Space-time trade-offs. In J. E. Savage, *Models of computation: Exploring the power of computing*. http://cs.brown.edu/~jes/book/pdfs/ModelsOfComputation.pdf

Shulman, C., & Bostrom, N. (2014). Embryo selection for cognitive enhancement: Curiosity or game-changer? *Global Policy, 5,* 85-92.

Sih, A., & Del Giudice, M. (2012). Linking behavioural syndromes and cognition: A behavioural ecology perspective. *Philosophical Transactions of the Royal Society of London B, 367,* 2762-2772.

Soukoreff, R. W., & MacKenzie, I. S. (2009). An informatic rationale for the speed-accuracy trade-off. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2890-2896). IEEE.

Spearman, C. E. (1927). *The abilities of man.* London: Macmillan.

Steinhäuser, C., Grunnet, M., & Carmignoto, G. (2016). Crucial role of astrocytes in temporal lobe epilepsy. *Neuroscience, 323,* 157-169.

Sterling, P., & Freed, M. (2007). How robust is a neural circuit? *Visual Neuroscience, 24,* 563-571.

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. Cambridge, MA: MIT Press.

Suffczynski, P., Kalitzin, S., & Da Silva, F. L. (2004). Dynamics of non-convulsive epileptic phenomena modeled by a bistable neuronal network. *Neuroscience, 126,* 467-484.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. New York, NY: Random House.

Tapia, D. H., Silva, A. X., Ballesteros, G. I., Figueroa, C. C., Niemeyer, H. M., & Ramírez, C. C. (2015). Differences in learning and memory of host plant features between specialist and generalist phytophagous insects. *Animal Behaviour, 106,* 1-10.

Todd, P. M., Gigerenzer, G., and the ABC Research Group (2012). *Ecological rationality: Intelligence in the world.* New York, NY: Oxford University Press.

Todd, P. M., & Miller, G. F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In G. Gigerenzer & P. M. Todd, & the ABC Research Group. (Eds.), *Simple heuristics that make us smart* (pp. 287–308). New York, NY: Oxford University Press.

Tomasi, D., Wang, G. J., & Volkow, N. D. (2013). Energetic cost of brain functional connectivity. *Proceedings of the National Academy of Sciences USA, 110,* 13642-13647.

Tops, M., Boksem, M. A. S., Luu, P., & Tucker, D. M. 1391 (2010). Brain substrates of behavioral programs associated with self-regulation. *Frontiers in Psychology, 1,* 1393 152.

Tosh, C. R., Krause, J., & Ruxton, G. D. (2009). Theoretical predictions strongly support decision accuracy as a major driver of ecological specialization. *Proceedings of the National Academy of Sciences USA, 106,* 5698-5702.

Tosh, C. R., Ruxton, G. D., Krause, J., & Franks, D. W. (2011). Experiments with humans indicate that decision accuracy drives the evolution of niche width. *Proceedings of the Royal Society of London B, 278,* 3504-3509.

Ungerleider, L. G., Doyon, J., & Karni, A. (2002). Imaging brain plasticity during motor skill learning. *Neurobiology of Learning and Memory, 78,* 553-564.

Vakulenko, S. A., & Radulescu, O. (2012). Flexible and robust networks. *Journal of Bioinformatics and Computational Biology, 10,* 1241011.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21,* 615-628.

van Noordwijk, A.J., & de Jong, G. (1986) Acquisition and allocation of resources: Their influence on variation in life history tactics. *American Naturalist, 128,* 137-142.

Wang, S. S. H., Shultz, J. R., Burish, M. J., Harrison, K. H., Hof, P. R., Towns, L. C., ... & Wyatt, K. D. (2008). Functional trade-offs in white matter axonal scaling. *Journal of Neuroscience, 28,* 4047-4056.

Warbrick, T., Rosenberg, J., & Shah, N. J. (2017). The relationship between BOLD fMRI response and the underlying white matter as measured by fractional anisotropy (FA): a systematic review. *NeuroImage, 153,* 369–381.

Wee, B., & Singer, M. C. (2007). Variation among individual butterflies along a generalist–specialist axis: no support for the 'neural constraint' hypothesis. *Ecological Entomology, 32,* 257-261.

Wenger, E., Brozzoli, C., Lindenberger, U., & Lövdén, M. (2017). Expansion and renormalization of human brain structure during skill acquisition. *Trends in Cognitive Sciences, 21,* 930-939.

Wisdom, T. N., Song, X., & Goldstone, R. L. (2013). Social learning strategies in networked groups. *Cognitive Science, 37,* 1383-1425.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B, 358,* 593–602.

Woodley, M. A. (2011). The cognitive differentiation-integration effort hypothesis: A synthesis between the fitness indicator and life history models of human intelligence. *Review of General Psychology, 15,* 228-245.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience, 16,* 487-497.

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nature Neuroscience, 15*, 528-536.

Zylberberg, J., Cafaro, J., Turner, M. H., Shea-Brown, E., & Rieke, F. (2016). Direction-selective circuits shape noise to ensure a precise population code. *Neuron, 89,* 369-383.