COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2014) 1–41

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12192

Mechanisms for Robust Cognition

Matthew M. Walsh, Kevin A. Gluck

Air Force Research Laboratory, Wright-Patterson AFB

Received 27 September 2013; received in revised form 14 March 2014; accepted 9 May 2014

Abstract

To function well in an unpredictable environment using unreliable components, a system must have a high degree of robustness. Robustness is fundamental to biological systems and is an objective in the design of engineered systems such as airplane engines and buildings. Cognitive systems, like biological and engineered systems, exist within variable environments. This raises the question, how do cognitive systems achieve similarly high degrees of robustness? The aim of this study was to identify a set of mechanisms that enhance robustness in cognitive systems. We identify three mechanisms that enhance robustness in biological and engineered systems: *system control*, *redundancy*, and *adaptability*. After surveying the psychological literature for evidence of these mechanisms, we provide simulations illustrating how each contributes to robust cognition in a different psychological domain: psychomotor vigilance, semantic memory, and strategy selection. These simulations highlight features of a mathematical approach for quantifying robustness, and they provide concrete examples of mechanisms for robust cognition.

Keywords: Robustness; System control; Redundancy; Adaptability; Cognitive systems; Computer simulation

1. Introduction

Robustness is the ability of a system to maintain its function despite perturbation. To survive in an unpredictable environment using unreliable components, a system must be robust. Robustness is fundamental to biological systems (Hammerstein, Hagen, Herz, & Herzel, 2006; Kitano, 2004). For example, chemical reactions that produce circadian rhythms must occur precisely despite widespread variation in ambient temperature (Hatakeyama & Kaneko, 2012). Robustness is also an objective in the design of engineered systems. For example, airplane engines must produce adequate thrust at different altitudes (Reinman, Ayer, Davan, Devore, & Finley, 2012) and buildings must withstand dynamic

Correspondence should be sent to Matthew M. Walsh, 711 HPW/RHAC – Cognitive Models and Agents Branch, 2620 Q Street, Building 852, Wright-Patterson AFB, OH 45433. E-mail: mmw188@gmail.com

forces (Starossek & Haberland, 2012). The capacity for robust responses to environmental variation is arguably a basic property of all complex, evolved systems (Hammerstein & Stevens, 2012), whether biological or artifactual.

Cognitive systems, like biological and engineered systems, exist within variable environments. This means that robustness is a key property of cognitive systems as well.[1] Indeed, neural circuits are frequently described as being *robust* (Cain, Barreiro, Shalden, & Shea-Brown, 2013; Noppeney, Friston, & Price, 2004) and capacities such as vision and memory are often called *robust* (Taatgen, Huss, Dickison, & Anderson, 2008; Wyatte, Curran, & O'Reilly, 2012). Moreover, the decision strategies, or heuristics, that people use are commonly said to be *robust* (Gigerenzer & Gaissmaier, 2011).

Despite an emerging consensus in the cognitive science community that robustness is a key property of cognition, there has been no serious attempt to identify the underlying mechanisms that allow cognitive systems to achieve this property. In this study, we take on precisely that objective—to identify a set of mechanisms for robust cognition.

Cognitive scientists make a distinction between representation and process. Although this is a valid distinction, in cognitive systems, "representation and process are inextricably interwoven, so that if our research teaches us something about one of them, it inevitably teaches us something about the other" (Simon, 1982, p. 334). Thus, our search for mechanisms includes both representations and processes that enhance robustness.

We begin by introducing robustness with examples from biology and engineering. These examples establish a multidisciplinary view of robustness and reveal a general set of mechanisms that also protect psychological functions against perturbations. We then describe a method for quantifying robustness (Walsh, Einstein, & Gluck, 2013). Such quantification is necessary to make precise statements about *degree* of robustness. We conclude by providing simulations illustrating how each mechanism supports cognition in a different psychological domain: psychomotor vigilance, semantic memory, and strategy selection. These simulations provide concrete examples of mechanisms for robust cognition.

Robustness is already an established theme in biology and engineering. In the following sections, we look to those fields for preliminary answers to two questions: Why is robustness important, and what mechanisms enhance robustness? The answers to these questions capture basic aspects of robustness that are relevant throughout cognitive science.

2. Why is robustness important?

Understanding why a system operates as it does require understanding the selective forces that determined its fitness, and thus shaped its evolution. For example, circadian rhythms are biological processes with endogenous, entrainable oscillations of about 24 h (Reppert & Weaver, 2002). These rhythms allow the organism to anticipate environmental change caused by the Earth's rotation, and so to synchronize behavioral and physiological processes to the appropriate time of day. The biochemical reactions underlying the circadian clock are dependent upon the organism's internal temperature, yet the clock operates over a range of temperature conditions. Understanding the circadian rhythm,

then, requires understanding how biochemical reactions produce an oscillation with a period of about 24 h, and how they do so as the temperature fluctuates (Hatakeyama & Kaneko, 2012).

A biological system is robust if it continues to function despite perturbation (Flack, Hammerstein, & Krakauer, 2012; Kitano, 2004). Biological perturbations can be genetic (e.g., mutations) or environmental (e.g., temperature change). These may threaten the survival of individuals in a species, or even the species as a whole. In this sense, evolution favors traits that confer robustness against biological perturbations.

Robustness is also an objective in structural engineering. Threats that may affect a structure (e.g., material fatigue, extreme weather, or explosions; Bontempi, Giuliani, & Gkoumas, 2007) are referred to collectively as its *exposure*. These are the "perturbations," a structural system is likely to encounter. Robust structures are resistant to progressive disproportionate collapse following exposure to such threats.

Finally, mechanical engineers seek to create robust systems. This is the motivation behind Design for Variation, a strategic initiative launched at Pratt and Whitney (Reinman et al., 2012). Central to this initiative is the idea that variation is a major stressor of fielded systems. For example, the effectiveness of turbine airfoil cooling features depends on several factors, including airfoil geometric dimensions, pressure, and temperature. By identifying the input variables that contribute most to uncertainty in airfoil temperature, engineers can focus on reducing variation in those inputs or enhancing robustness of the design against those sources of variation.

Common across these domains is the idea that robustness is the ability of a system to maintain its function despite perturbation. The meaning of function is context-specific. The function of a biological system may be to survive and reproduce, whereas the function of a structural system may be to provide shelter or support. Likewise, the meaning of perturbation is context-specific. Perturbations include genetic mutations (Wagner, 2005), the variable temperatures over which the circadian clock or a turbine airfoil must operate (Hatakeyama & Kaneko, 2012; Reinman et al., 2012), or changes in load distribution following loss of structural elements (Starossek & Haberland, 2012). Robustness is important because it allows systems to function across the variable environments they encounter (Gluck et al., 2012).

3. What mechanisms enhance robustness?

A shared set of mechanisms enhances robustness across domains (Table 1).[2] The fitness advantages conferred by these mechanisms explain why they reoccur as products of natural and artifactual design processes.

3.1. System control: Using a measure of oneself or one's output to adjust control

System control includes negative feedback loops, positive feedback loops, or other regulatory feedback loops that allow the system to remain in one state or to move among

Table 1
Mechanisms for robustness

Mechanism	Biology	Engineering	Cognitive Science
System control	Feedback loops, integral control	Closed-loop control	Feedback control, conflict detection, recurrent connectivity
Redundancy	Parallel metabolic pathways	Alternate load paths	Redundant and diverse neural systems, distributed representations
Adaptability	Inducible defenses	Adaptive controllers	Strategy selection, mixture-of-experts

stable states. For example, bacterial chemotaxis, the process by which bacteria track chemical gradients, is robust against variations in the strength of extracellular attractant signals and intracellular protein concentrations (Barkai & Leibler, 1997). Bacteria use a form of negative feedback control called integral control to adapt to these sources of external and internal variation (Yi, Huang, Simon, & Doyle, 2000).

Control theorists distinguish between closed-loop systems, which use a measure of their output to modify control signals, and open-loop systems, which do not (Goodwin, Graebe, & Salgado, 2001). By monitoring the system's output, closed-loop controllers can compensate for external disturbances acting on the system. Closed-loop control, also called feedback control, is commonly used to manage devices that must operate in variable conditions (e.g., thermostats).

3.2. Redundancy: Having multiple components to achieve the same function

If homogenous components exist, one can replace another following failure. Such pure redundancy, though common in engineered systems, is rare in nature. More typically, heterogeneous components support the same function but in different ways. This form of redundancy is also called *redundancy and diversity* (Kitano, 2004), or *degeneracy* (Edelman & Gally, 2001; Friston & Price, 2003). For example, oxidative phosphorylation and glycolysis both produce adenosine triphosphate (ATP; Berg, Tymoczko, & Stryer, 2002), which is crucial for intracellular metabolic processes. Oxidative phosphorylation, the primary source of ATP, requires oxygen, whereas glycolysis does not.

Redundancy is also present in engineered systems. For instance, building codes call for alternate load paths in structural designs (Canisius, 2011; NISTIR 7396, 2007). These load paths can redistribute forces originally carried by failed components, arresting incipient collapse. Likewise, automatic flight control systems contain multiple computers. Each performs the same function in a different way to ensure redundancy and diversity (Pratt, 2000).

3.3. Adaptability: Switching into different modes of operation to maintain performance

Robustness does not mean that genetic and environmental perturbations do not affect how a system behaves. Perturbations may force a system to enter a new operational mode

to maintain functionality (Hagen & Hammerstein, 2005). For example, water fleas can grow a helmet-like structure for protection against predators. Because the helmet is costly to construct, fleas only grow it upon detecting traces of the predator (Agrawal, Laforsch, & Tollrian, 1999). This inducible defense is a dramatic example of an environment-specific, phenotypic adaptation.

Adaptability also appears in engineered systems in the form of adaptive control (Ioannou & Sun, 1996). Because control solutions are highly specialized, a solution that performs well in one context may perform poorly in others. For example, the dynamic properties of the CH-47 helicopter change as a function of its horizontal and vertical flight velocities (Sastry & Bodson, 1989). Airspeed sensors monitor these variables to select which of 90 different controllers to activate as the CH-47 operates.

Adaptability relates to the first and second mechanisms for robust performance. To be adaptable, a system must have multiple ways to achieve a function (i.e., redundancy). The system must also have a sensor to monitor itself and the environment, and a controller to select among redundant components (i.e., system control). The controller must select components that are appropriate given the state of the system and environment. This may require hardcoding knowledge into the system, or incorporating learning algorithms into its design.

4. Robustness of the cognitive system

Like biological and engineered systems, the cognitive system must function within variable environments. To the extent that the mechanisms we identified are general, these same mechanisms may enhance cognitive robustness. We now review evidence that this is, in fact, the case.

4.1. System control

One distinction in the motor control literature is between feedforward and feedback control (Desmurget & Grafton, 2000; Jordan, 1996; Kawato, 1999). In feedforward control, a sequence of muscle activations that produce the desired state is determined before movement onset. Once a command signal is issued, the action is performed without modification. In feedback control, the system's current state is compared against the desired state. As an action is performed, positional errors are used to generate corrective commands. To generate a corrective signal, the motor system must estimate its current state. This is done using sensory (e.g., visual and prioceptive) and non-sensory (e.g., efferent motor commands) information (Desmurget & Grafton, 2000; Todorov, 2004). Feedback control enhances robustness by allowing the motor system to make adjustments when external forces are applied to the active effector, and by allowing the motor system to correct movements that were incorrectly specified during planning.

More generally, system control is the ability of a system to adjust its dynamics while monitoring itself, and not just its output. Neuroimaging and electrophysiology studies have reported anterior cingulate cortex (ACC) activation in tasks that involve (a) overriding prepotent responses; (b) selecting among equally correct responses; or (c) error commission. In all of these tasks, multiple incompatible responses are simultaneously active. Based on this observation, Botvinick, Braver, Barch, Carter, and Cohen (2001) proposed that the ACC monitors for response conflict. Upon detecting conflict, the ACC engages areas such as the dorsolateral prefrontal cortex, which then increase attentional control. This strengthens stimulus inputs and reduces response conflict. By this view, the ACC contributes to cognitive robustness by monitoring for response conflict and signaling for greater control as needed.

Finally, system control in the form of positive feedback loops enhances cognitive robustness by amplifying internal representations of degraded stimulus inputs. The hippocampus, and the medial temporal lobe more generally, support episodic knowledge (Squire, 1987). When cues that comprise a memory are encountered, the hippocampus activates the corresponding, internally stored episode. Importantly, people can retrieve episodic memories given incomplete and degraded inputs. To account for this ability, Marr (1971) developed an autoassociative model of the hippocampus. Autoassociative networks are trained to map input patterns to identical output patterns (Kohonen, 1984). The output is sent back to the network as an input signal through recurrent connections. When the network is presented with an incomplete pattern, activation from the recurrent collaterals iterates through the network and gradually completes the missing parts of the pattern (i.e., pattern completion).

The recurrent connectivity of Marr's autoassociative network, and of other hippocampal models it inspired (Gluck & Myers, 1997), forms a positive feedback loop. The activation pattern corresponding to a stored episode constitutes a stable state. Recurrent connections allow the network to move among states based on the input. Consequently, the output of the network is robust against variation in the completeness and quality of inputs. Recurrent networks have also been used to model working memory that is robust against temporal delay (O'Reilly & Frank, 2006), and object recognition that is robust against visual occlusion and image degradation (Wyatte et al., 2012).

4.2. Redundancy

A remarkable feature of the human brain is its ability to maintain and recover cognitive functions following focal cortical damage. This challenges the notion of a one-to-one mapping between structure and function, and it suggests that each function can be accomplished by many structures (Just & Varma, 2007; Lashley, 1950; Price & Friston, 2002). These structures are not mere replicates. Different structures perform the same function in different ways (Edelman & Gally, 2001; Friston & Price, 2003). The existence of separate information-processing mechanisms that yield consistent outputs is recognized in cognitive neuroscience. For example, regular words can be read using phonological or semantic processes (Seidenberg & McClelland, 1989). Regular word reading, therefore, is

left relatively intact following circumscribed damage to neural structures associated with one of the pathways.

Functional-magnetic resonance imaging experiments and patient studies are informative with respect to this issue (Noppeney et al., 2004; Price & Friston, 2002). Functional imaging experiments can identify the set of regions sufficient for performing a task, whereas patient studies can determine which of those regions are necessary. For example, when asked to match words on the basis of their meaning, neurotypical participants activate a distributed semantic-retrieval network (Mummery et al., 1999). However, focal lesions encompassing each of the regions that comprise the network do not dramatically alter performance. Thus, although the set of regions sufficient for performing the task is large, few are necessary (Price & Friston, 2002).

As another example, multiple neural systems regulate instrumental responding. The emerging view from behavioral and physiological studies is that two forms of control, goal-directed and habitual, coexist as complementary mechanisms for action selection (Balleine & O'Doherty, 2010; Daw, Niv, & Dayan, 2005). Goal-directed control is dominant early in training, and habitual control after extended training. If neural structures responsible for goal-directed control are lesioned early in training, however, habitual control is expressed immediately (Yin, Ostlund, Knowlton, & Balleine, 2005). Alternatively, if neural structures responsible for habitual control are lesioned after extended training, goal-directed control is re-expressed (Yin, Knowlton, & Balleine, 2004). The distinction between goal-directed and habitual control is surprisingly general; similar definitions arise in artificial intelligence and in cognitive modeling (Walsh & Anderson, 2014). In all cases, the existence of multiple controllers confers robustness against damage to one.

Lastly, robust internal representations depend on redundancy. The graceful degradation that follows focal cortical damage is an emergent property of some connectionist models of cognitive phenomena (Hinton & Sejnowski, 1986). Processing in connectionist models occurs through the propagation of activation values through networks of units. Knowledge is stored as the set of connection weights among units, and each concept (e.g., "bird") is represented by a pattern of activation distributed across many units. Distributed representations arise from the combined effect of many units, no one of which is critical to the realization of the representation. Consequently, when units are removed or noise is added to connection weights, representations, though degraded, remain.

4.3 Adaptability

Psychological research has inspired the view that the mind contains a toolbox of decision strategies, or heuristics (Gigerenzer & Gaissmaier, 2011). When faced with choices, people can, and do, use a variety of heuristics. Such variability enhances robustness if the individual can choose adaptively among strategies. Many models of human problem solving include a strategy selection process (Payne, Bettman, & Johnson, 1988; Rieskamp & Otto, 2006; Siegler, Adolph, & Lemaire, 1996; Walsh & Anderson, 2009). Upon viewing a problem, the individual evaluates the applicability of different strategies. This evalua-

tion is informed by the history of strategy use: What has worked in the past will likely work again. This evaluation is also informed by the current context: Different strategies are applicable to problems with different features. Based on past experiences and the current context, the individual attempts to identify the strategy that will maximize performance. By selecting strategies that are tuned to the environment, and are thus ecologically rational (Marewski & Schooler, 2011), the individual can perform well even as features of the problem or decision scenario change.

Adaptability also extends to the motor system, which may switch among multiple pairs of forward and inverse models to control behavior (Wolpert & Kawato, 1998). The idea that we use internal models to plan actions is an important concept in motor control (Brown & Rosenbaum, 2002; Jordan, 1996; Kawato, 1999; Wolpert, 1997). Forward models predict the sensory consequences of motor commands, and inverse models identify motor commands that will produce the desired output. Both types of models mimic the input/output characteristics of the motor apparatus. But the range of objects and environments that the sensorimotor system interacts with is too great to allow one internal model to generalize to all possible scenarios (Kawato, 1999; Wolpert, 1997). For example, the kinematic and dynamic properties of a brush differ from those of a racquet. Consequently, the same internal model cannot be used to plan brush and racquet strokes.

The mixture-of-experts architecture provides a solution to this problem (Jacobs, Jordan, Nowlan, & Hinton, 1991). The mixture-of-experts architecture consists of a set of expert modules, each of which is specialized for a different subtask. A gating module combines the outputs of all expert modules. In the *multiple paired forward-inverse model* (Wolpert & Kawato, 1998), an instance of this type of architecture, expert modules contain corresponding forward and inverse models. A gating module transfers control among experts based on how well their forward models predict incoming sensorimotor signals, and based on contextual information from sensory inputs. By transferring control among expert modules, the multiple paired forward–inverse model maintains performance even as kinematic and dynamic properties of the motor apparatus change.

5. Quantification of robustness

Having identified a collection of mechanisms that enhance robustness—system control, redundancy, and adaptability—and having established their relevance to cognitive science, we turn our attention to a formal quantification of robustness. We conceptualize robustness as *the extent to which a system is able to maintain its function when some aspect of the system is subject to perturbation* (Gluck et al., 2012). This serves to clarify our definition of robustness, but it does not itself advance a methodology for measuring degree of robustness and selecting based on that property. To that end, we now summarize a methodological operationalization of robustness (Walsh et al., 2013).

To quantify robustness, one must specify the system's function. Functionality may be defined by a single performance goal (e.g., to respond accurately), or by multiple performance goals (e.g., to respond accurately and quickly). One must also specify the

perturbations the system may encounter. We treat the environment as a multidimensional space, E. Dimensions can be internal to the individual (e.g., degree of fatigue) or external to the individual (e.g., extent of object occlusion). Each scenario the individual faces, x, constitutes a point in space defined by values along the dimensions that comprise the environment ($x \in E$). Perturbations involve altering values along the dimensions to move the individual from one scenario to another.

After defining a system's function and identifying perturbations the system may encounter, one can quantify robustness. At a high level, this involves interpreting performance metrics (i.e., reaction times and response accuracy) based on contextual factors (i.e., scenarios and tolerance). Our approach entails three steps: calculate functionality, assess robustness, and measure stability.

5.1. Step 1: Calculate functionality

First, we ask, did the system achieve its function in an isolated scenario? Functionality in scenario x is defined as

$$\text{Functionality}(x) = \left(S(x) = \frac{F(x)}{T} \right) \tag{1}$$

Success [$S(x)$] is the proportion of time the system achieved its function, and failure [$F(x)$] is the proportion of time the system did not [$S(x) + F(x) = 1$]. Tolerance ($T > 0.0$) depends on the significance of failure. If errors are consequential, tolerance must be low. If errors are inconsequential, tolerance may be higher. Functionality is bound by 1.0 for when $S(x) = 1.0$, and by $-1/T$ for when $F(x) = 1.0$.

Tolerance relates to risk-based criteria (i.e., maintain below 3.4 failures per million opportunities; $0.0 < \text{risk} < 1.0$). The interpretation of functionality is straightforward when tolerance is set to the ratio of acceptable risk to one minus acceptable risk, $T = \text{risk}/(1-\text{risk})$. When tolerance is set in this way, negative functionality denotes that the system violates acceptable risk, and positive functionality denotes that the system remains within acceptable risk. In applied settings, acceptable risk and tolerance can be derived from established or expressed performance standards (Walsh et al., 2013). Tolerance can also be determined with a risk matrix or an economic analysis of total costs, tools that have been used to identify target levels of reliability in engineering and medical decision making (Huaco, Bowders, & Loehr, 2012; Weinstein & Fineberg, 1980).

Responses cannot always be classified simply as correct or incorrect. When multiple types of correct responses and errors are possible, different tolerances may be assigned to each response category,

$$\text{Functionality}(x) = \left([S_1(x) + S_2(x) \ldots + S_m(x)] - \left[\frac{F_1(x)}{T_1} + \frac{F_2(x)}{T_2} + \ldots \frac{F_n(x)}{T_n} \right] \right) \tag{2}$$

$S_{1...m}$ are the proportions of successful responses in each of m categories, and $F_{1...n}$ are the proportions of unsuccessful responses in each of n categories. $T_{1...n}$ are the specific tolerances for each type of failure.

To facilitate interpretation, functionality scores are normalized according to the strictest tolerance,

$$\text{Functionality}_{\text{norm}}(x) = \frac{\min(T_{1...n}) \cdot \text{Functionality}(x) + 1}{\min(T_1...n) + 1} \tag{3}$$

In Eq. (3), $\min(T_{1...n})$ is the minimum tolerance from the n types of possible failure. Following normalization, all functionality scores fall between zero (minimally functional) and one (maximally functional). Prior to normalization, positive functionality denotes that the system remains within allowable risk. Following normalization, functionality greater than $1/(\min(T_{1...n})+1)$ denotes that the system remains within allowable risk.

5.2. Step 2: Assess robustness

Next, we ask, to what extent did the system achieve its function across all scenarios formed by combinations of the environment dimensions? Robustness is defined as

$$\text{Robustness} = \int \text{Functionality}_{\text{norm}}(x) \cdot \text{Probability}(x)dx \tag{4}$$

Probability(x) describes the probability of encountering that scenario. In the absence of prior information, all scenarios are treated as equally likely. Because normalized functionality is bounded by zero and one, and because the probability distribution of scenarios sums to one, robustness is bound by zero (minimally robust) and one (maximally robust).

5.3. Step 3: Measure stability

Lastly, we ask, how much did functionality vary across scenarios? Stability is defined as

$$\text{Stability} = 1 - 2 \cdot \sigma_{\text{Functionality}_{\text{norm}}} \tag{5}$$

The range of the standard deviation among normalized functionality scores is 0.0–0.5. We multiply the standard deviation by two to place stability over the range 0.0–1.0, and we treat stability as a decreasing function of the standard deviation because stability is inversely related to the degree of variability among functionality scores. Stability is bound by zero (minimally stable) and one (maximally stable).

Robustness and stability are orthogonal measures. Robustness answers the question, "Does the system perform well on average?" Stability answers the slightly different question, "Is the performance of the system insensitive to perturbation?" Both are desirable properties of a system.

5.4. Implementation

Having described a method for quantifying robustness, we now use it to demonstrate how each mechanism contributes to robust cognition in a different psychological domain: psychomotor vigilance, semantic memory, and strategy selection. In each example, we use simulations to estimate functionality in the scenarios created by applying perturbations. For instance, the first simulation, psychomotor vigilance, involves a single perturbation: time awake. We measure the functionality of two models of psychomotor vigilance in 45 scenarios created by varying time awake from 0 to 88 h in 2-h increments. The models predict response time distributions for each scenario. These distributions serve as the basis for calculating the proportions of trials resulting in success (S in Eq. 2) and failure (F in Eq. 2), which are used to measure functionality.

Robustness and stability describe functionality over all scenarios comprising the environment. Rather than calculating these values analytically (Eqs. 4 and 5), we estimate them using numerical simulations over the subset of enumerated scenarios. To quantify robustness and stability, we generate samples. Each sample contains a functionality score from all of the 45 scenarios. The mean of functionality within a sample is used to estimate robustness, and the standard deviation of functionality within a sample is used to estimate stability. Because simulations are stochastic, functionality in a scenario varies for each model run. By estimating robustness and stability repeatedly using a large number of samples, we can quantify uncertainty in these estimates.

6. Simulation studies

In each of the following simulation studies, we calculate the robustness and stability of two or more cognitive systems. Comparative analysis of multiple systems is a straightforward application of the quantification described in the previous section. Functionality, and therefore robustness, depends on tolerances assigned to the different types of failure. Because tolerance is set by context, differences among the robustness of cognitive systems within the same context are meaningful. The robustness value of a single system is also meaningful. Normalized functionality greater than $1/(\min(T_{1\ldots n})+1)$ indicates that cognition is functioning within allowable risk levels. This value provides a criterion against which to determine whether a single system violates allowable risk, how close the system comes to violating allowable risk, or how often the system violates allowable risk.

7. Simulation 1. System control in psychomotor vigilance

Vigilance is the ability to maintain focused attention for extended periods of time (Davies & Parasuraman, 1982). The systematic study of vigilance was first motivated by the question of why radar operators sometimes missed weak signals indicating the presence of enemy submarines (Mackworth, 1948). Vigilance remains an important topic.

Advances in automated technology have shifted the role of human operators from active controllers to supervisors who must react when problems arise (Warm, Parasuraman, & Matthews, 2008). Consequently, much research has gone toward understanding vigilance and factors that affect it.

One of these factors is fatigue. An extensive literature documents the deleterious effects of fatigue on psychological functioning (for a review, see Durmer & Dinges, 2005). As part of the effort to understand these effects, researchers have developed biomathematical models of fatigue. Nearly all are based on the canonical two-process theory (Borbély & Achermann, 1999): Circadian processes produce variations in alertness over the course of a single day, whereas sleep homeostatic processes produce continuously declining alertness with time awake.

Importantly, these models and the experiments that motivate them do not show that sleep loss simply suppresses neurobehavioral functions. Rather, sleep loss destabilizes performance (Doran, van Dongen, & Dinges, 2001). The circadian component of fatigue has been linked with the suprachiasmatic nucleus, and the sleep homeostatic component with the accumulation of adenosine—a byproduct of ATP—and sleep regulatory substances. Consolidation of circadian and homeostatic processes occurs in subcortical sleep regulatory circuits, which send ascending projections to the thalamus and cerebral cortex (for a review, see Saper, Scammell, & Lu, 2005). The behavioral variability people display when fatigued is thought to arise from the interaction between the drive for sleep, and compensatory behavioral and neurophysiological responses to remain alert (Drummond, Brown, Salamat, & Gillin, 2004). In other words, the brain monitors for and increases effort upon detecting the onset of fatigue. This *system control mechanism* enhances the robustness of the cognitive system against fatigue.

7.1. The psychomotor vigilance task (PVT)

The PVT is a sustained attention task used to measure alertness (Dinges & Powell, 1985). Participants monitor a display for the presentation of a stimulus. When the stimulus appears, they respond. *False starts* are trials in which a response is made before or within 150 ms of stimulus presentation, and *sleep attacks* are trials in which no response is made within 30s of the stimulus. In an experiment by Doran, van Dongen, and Dinges (2001), participants remained awake for 88 h and completed the PVT every 2 h. With increasing time awake, the number of false starts increased, the number of sleep attacks increased, and alert responses slowed. These outcomes of the PVT are highly replicable and predictable (Lim & Dinges, 2008). But despite these performance decrements, participants retain remarkable ability to perform the PVT even after extended periods of wakefulness.

7.2. Task model

Gunzelmann, Gross, Gluck, and Dinges (2009) modeled the effects of sleep deprivation in the PVT using Adaptive Control of Thought–Rational (ACT-R; Anderson, 2007). ACT-R

is a production system with a set of specialized information-processing modules. Buffers connect these modules with a central procedural module. Procedural knowledge is represented in the form of production rules. Each rule has a set of conditions that must be met for it to be selected, and a set of actions that modify the external state of the world and the internal state of the architecture. The temporal dynamics of cognition unfold across a series of production cycles. In each cycle, conditions for different productions are compared against the contents of the buffers, and the production with greatest utility is selected and enacted. The resulting state of the world and of the architecture serves as the starting point for the next production cycle.

Gunzelmann et al. (2009) included four productions in their model of the PVT: *wait* for the stimulus to appear, *attend* to the stimulus once it appears, *respond* after attending to the stimulus, and *respond randomly*. The fourth production, though rarely chosen because of its low utility, caused false starts. Subsequent to that model, ACT-R underwent several major changes, including the addition of production partial matching. With production partial matching, productions whose conditions are not perfectly met remain eligible for selection, but their utility values are penalized. Because of production partial matching, *respond* can be selected at any time. This obviates the need for the *respond randomly* production: When *respond* is selected before the stimulus appears, a false start occurs. But because *respond* is subject to the mismatch penalty when its conditions are not perfectly met, this happens infrequently.

In a revised ACT-R model of the PVT (Walsh, Gunzelmann, & Van Dongen, 2014), instantaneous utility is calculated as

$$U_i' = \mathrm{FP} \cdot (U_i - \mathrm{MMP}_i) + \varepsilon_i \tag{6}$$

U_i is the stored utility of production i, MMP_i is the mismatch penalty, and ε_i is logistically distributed noise. The alertness value FP is derived from a biomathematical model of fatigue (McCauley et al., 2013). FP varies with the circadian rhythm and decreases with time awake. Consequently, as fatigue increases, the utilities of all productions approach zero.

In each production cycle, the rule with highest utility is enacted if its utility exceeds a threshold,

$$\mathrm{Choice} = \max(U_i'), \mathrm{if}\ \ \max(U_i') > \mathrm{FT} \cdot \mathrm{Threshold} \tag{7}$$

If no production's utility exceeds the threshold, a micro-lapse occurs: The model becomes briefly inactive before searching again for a rule with sufficiently high utility to enact. Because production utilities decrease with fatigue, it becomes unlikely that the value of any production will exceed the threshold.

To offset this effect, Walsh et al. (2014) dynamically adjusted the threshold in accord with the moderating variable, FT.[3] As with the production utility moderating variable FP, the threshold moderating variable FT is derived from a biomathematical model of fatigue (McCauley et al., 2013). Both decrease with time awake, but FT does so more gradually. This is a computational instantiation of the neurophysiological

system control mechanism described earlier. Decreasing the utility threshold with time awake increases the probability that a production will remain eligible for selection. Thus, behavior of the integrated model depends on the interaction between a primary response—the decreased production utilities caused by the onset of fatigue, and a compensatory response—the decreased utility threshold following the detection of fatigue.

7.3. Simulation 1A: PVT

We tested versions of the PVT model with and without compensatory adjustment of the utility threshold (FT in Eq. 7). This comparative manipulation of the presence or absence of a system control mechanism enables precise assessment of the relative degree of cognitive system robustness. The simulation involved a single perturbation, time awake. For each model, we simulated 100 participants at times ranging from 0 to 88 h in 2-h increments (based on the experiment by Doran, van Dongen, & Dinges, 2001). For this, the preliminary simulation, we only present data on the models' response time distributions. These distributions serve as the basis for computing robustness and stability in the next simulation.

Fig. 1 shows the cumulative reaction time distributions averaged over the baseline period and over each subsequent day of the experiment. In the compensatory model, the proportion of trials without responses (i.e., sleep attacks) increased from .001 ± 0.002 SD at baseline to 0.023 ± 0.002 SD on Day 3. The size of the performance decrement in the non-compensatory model was nearly 10 times greater; the proportion of trials without responses increased from 0.001 ± 0.002 SD at baseline to 0.236 ± 0.007 SD on Day 3. In addition, alert responses in the compensatory model slowed from 0.50 s ± 0.03 SD at baseline to 1.56 s ± 0.07 SD on Day 3. Alert responses in the non-compensatory model slowed far more from 0.50 s ± 0.03 SD at baseline to 4.25 s ± 0.15 SD on Day 3. Finally, the compensatory model, like experiment participants, committed fewer false starts at baseline (0.069 ± 0.005 SD) than on Day 3 (0.143 ± 0.006 SD). The non-compensatory model actually showed a slight decrease in the proportion of trials with false starts (Baseline = 0.068 ± 0.005 SD; Day 3 = 0.061 ± 0.004 SD).

7.4. Simulation 1B: Driver model

The compensatory PVT model appeared more robust against fatigue. To confirm this analytically, we studied the two models in situ. Cognitive processes measured by the PVT underlie complex skills (Lim & Dinges, 2008). For instance, while driving, one must monitor traffic and decelerate upon detecting brake lights. We modified the PVT model to predict performance in a simplified driving scenario. To do so, we set the durations of motor planning (200 ms) and execution (500 ms) to values previously established for a braking motion (Salvucci, 2006). We adopted a driving scenario described in collision avoidance studies (Gray, 2011; Scott & Gray, 2008). In brief, the scenario involves trailing a lead car by 2 s while traveling at 24.5 m/s (55 mph). Given

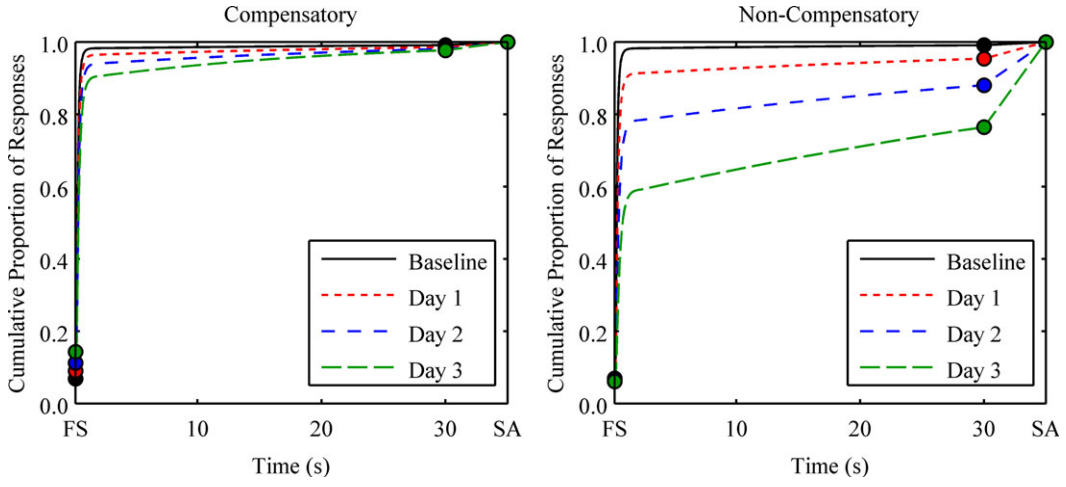

Fig. 1. Psychomotor vigilance task simulations. Cumulative distribution of reaction times averaged over Baseline (hours 0–16), Day 1 (hours 17–40), Day 2 (hours 41–64), and Day 3 (hours 65–88). Time bins are from 150 ms to 30 s in 1 ms increments. "FS" denotes trials with false starts (<150 ms), and "SA" denotes trials with sleep attacks (>30 s).

the deceleration of the lead car (−6 m/s), and the maximum deceleration of one's own car (−8.75 m/s), one must respond within 2.66 s of the lead car's braking to avoid collision.

We simulated the driving performance of 100 participants at times ranging from 0 to 88 h awake in 2-h increments. The resulting reaction time distributions resembled those in Fig. 1 but were shifted to the right due to the longer durations of the motor planning and execution times for the braking motion. From these distributions, we computed the proportions of trials with responses before 150 ms of the signal (*false starts*; F_1 in Eq. 2), within 2.66 s of the signal (*alert responses*; S in Eq. 2), and after 2.66 s of the signal (*collisions*; F_2 in Eq. 2). We assigned different tolerances to false starts ($T_1 = 0.50$) and collisions ($T_2 = 0.025$) because the consequences of collision exceed the consequences of premature braking. After normalizing functionality scores (Eq. 3), we computed robustness (Eq. 4) and stability (Eq. 5) over the range of the fatigue perturbation, using samples from the 100 simulations.

In both models, functionality varied over a 24-h period reflecting the contribution of the circadian component to FP, and functionality decreased over 88 h reflecting the contribution of the sleep homeostatic process to FP (Fig. 2). The data contain three other important effects. First, the models had nearly identical functionality at baseline (Table 2; $t(198) = 1.81$, $p > .05$, $d = .25$). Second, functionality of the non-compensatory model dropped sharply with time awake, whereas functionality of the compensatory model decreased gradually. This resulted in a greater value of robustness for the compensatory model (0.940 ± 0.002 SD vs. 0.802 ± 0.003 SD). Third, functionality of the non-compensatory model oscillated sharply within each day, whereas functionality of the compensatory model oscillated only slightly. This, paired with the smaller change in

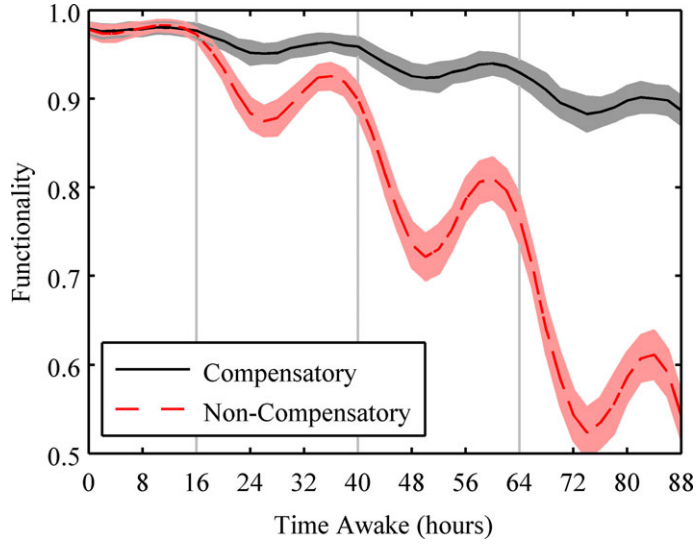

Fig. 2. Driver model simulations. Functionality of compensatory and non-compensatory models as time awake increases from 0 to 88 h in 2-h increments. Shaded regions show ±1 standard deviation (SD).

Table 2
Mean functionality of compensatory and non-compensatory driving models over four test periods, with standard deviations in parentheses

Model	Baseline	Day 1	Day 2	Day 3
Compensatory	0.978 (.002)	0.960 (.003)	0.934 (.004)	0.896 (.005)
Non-Compensatory	0.978 (.002)	0.909 (.005)	0.779 (.006)	0.586 (.008)

functionality over 88 h, resulted in a greater value of stability for the compensatory model (0.932 ± 0.004 SD vs. 0.690 ± 0.007 SD). Though the analysis only extends to 88 h awake, these effects hold over any interval beyond baseline.

Based on the minimum tolerance for collisions ($T_2 = 0.025$), the models perform within allowable risk when normalized functionality exceeds 0.9756. Neither model remained within allowable risk over 88 h of sleep deprivation (Table 2), and nor should they. Although the compensatory model is more robust than the non-compensatory model, it is not sufficiently robust to ensure safe driving after 88 h awake. A further step could be to use the robustness analysis to determine thresholds for factors such as maximum time awake, vehicle speed, and trailing distance to maintain allowable risk. We save such analyses for future applications of this methodology.

7.5. Discussion

ACT-R entails theoretical commitments linking architectural components with specific brain regions (Borst & Anderson, 2013). Most relevant here is the idea that production

rules are instantiated in the basal ganglia (Anderson, 2007). The main input structure of the basal ganglia, the striatum, sends predominantly inhibitory projections to the pallidum, which in turn inhibits cells in the thalamus. This creates a "winner-lose-all" situation in which active striatal neurons inhibit action representations in the pallidum, which in turn release the corresponding representations from inhibition in the thalamus. The winning action is passed to motor cortex only when striatal activation exceeds thalamic inhibition.

The utility threshold and the compensatory response to fatigue have been associated with the thalamus (Gunzelmann et al., 2009). In support of this view, task-related thalamic activation during alert performance increases with sleep deprivation (Chee & Choo, 2004; Habeck et al., 2004; Portas et al., 1998). In addition, behavioral lapses are accompanied by decreased activity in the thalamus (Chee et al., 2008; Paus et al., 1997). Thus, both the computational architecture and the neural data point to a compensatory system control mechanism in the thalamus for robust cognition even when fatigued.

The behavioral data also point to a compensatory system control mechanism for guarding performance against fatigue. PVT studies consistently show that with increasing time awake, alert responses slow and sleep attacks occur more frequently (Lim & Dinges, 2008). Both models produce these effects, but the non-compensatory model does so to a far greater extent than experiment participants do. In other words, people display a degree of robustness akin to the compensatory model.[4] Furthermore, PVT studies show that with increasing time awake, participants commit more false starts. This was the original reason for including compensatory adjustment of the utility threshold in the ACT-R model—decreasing the utility threshold offsets the effect of the mismatch penalty on the *respond* production, causing more false starts (Fig. 1). But as these simulations show, decreasing the utility threshold is vital to compensate for the effect of fatigue on utility values to limit the number of lapses. This interplay between a primary and compensatory response to fatigue is not unique to the ACT-R model. The effects of fatigue on psychomotor vigilance have been captured in a similar way in a diffusion model of the PVT (Walsh et al., 2014).

8. Simulation 2: Redundancy in semantic memory representations

How do we know what attributes an item has? Once acquired, how is this knowledge protected against the inevitable loss of neocortical neurons that occurs over the lifespan (Pakkenberg & Gundersen, 1997)? The connectionist framework provides a way to deal with these issues (Rogers & McClelland, 2004). In connectionist models, cognitive processes arise from interactions among simple, neuron-like units. Knowledge is stored in the strength of connections among units and is acquired gradually from experience.

Early models treated semantic knowledge, knowledge about items and their attributes, as a set of concepts (e.g., *robin*) and propositions (e.g., *can fly*). To determine whether a proposition was true of a concept, one accessed the concept to see whether the proposi-

tion was stored there (Collins & Quillian, 1969). In some connectionist models of semantic knowledge, concepts and propositions are not stored as such. Rather, concepts are represented by distributed patterns of activation across several units (Rogers & McClelland, 2004). The type of representation used in these models, called a distributed representation, is inherently redundant. Concepts are not represented by the activation of one unit, but by the activation of several units at once. This form of *redundancy* enhances the robustness of high-level representations against the loss of low-level units.

8.1. Task model

Fig. 3 shows a semantic network built by Rogers and McClelland (2004; cf. Rumelhart & Todd, 1993). The network has a feedforward structure; activation flows in one direction from item (e.g., *robin*) and relation units (e.g., *can*), through hidden layers, and to an output layer that contains units for completing propositions (e.g., *grow, move,* and *fly*). Connection weights among units are initially set to small random values so that inputs produce weak, undifferentiated outputs. The network is presented with examples during training. For instance, the inputs *robin* and *can* are set to one. Activation passes through the hidden layers and to output units. Activation of each output unit is compared against its target value (*grow*, *move*, and *fly* should equal one, and all other outputs should equal zero). The resulting error is used to adjust the strength of connections among units to reduce differences between actual and target outputs.

The model utilizes localist input representations; each item is signified by activation of one unit in the item layer. All units in the item layer connect with all units in the representation layer. This allows the network to re-encode localist inputs in a distributed fashion. During training, the distributed patterns of activation across representation units come to resemble a virtual semantic hierarchy. The more similar two items are, the more similar the patterns of activation they produce. Over the course of training, distributed representations distinguish first between superordinate categories (i.e., *plants/animals*), next among intermediate categories (i.e., *trees/plants* and *birds/fish*), and last among individual items (i.e., *oak/pine*, *rose/daisy, robin/canary*, and *salmon/sunfish*).

We examined the effects of three forms of redundancy on the robustness of the semantic network against simulated neural damage. The first form, representational redundancy, is the degree to which an attribute is shared by related concepts. For example, the attribute *can move* is shared by all concepts in the superordinate category of animals, whereas the attribute *can fly* is shared only by concepts in the intermediate category of birds. Thus, the attribute *can fly* is less redundant because it is shared by fewer related concepts. The attribute *can sing* is even less redundant because it is true only of an individual item, *canary*.

The second form of redundancy is architectural. The choice of how many units to include in the representation layer of the semantic network is subjective. Although earlier versions of this model include eight, Rumelhart and Todd (1993) found that a network

with fewer representation units performed equally well. This raises the question of whether there is an advantage to having more than the smallest number of representation units necessary to achieve mastery.

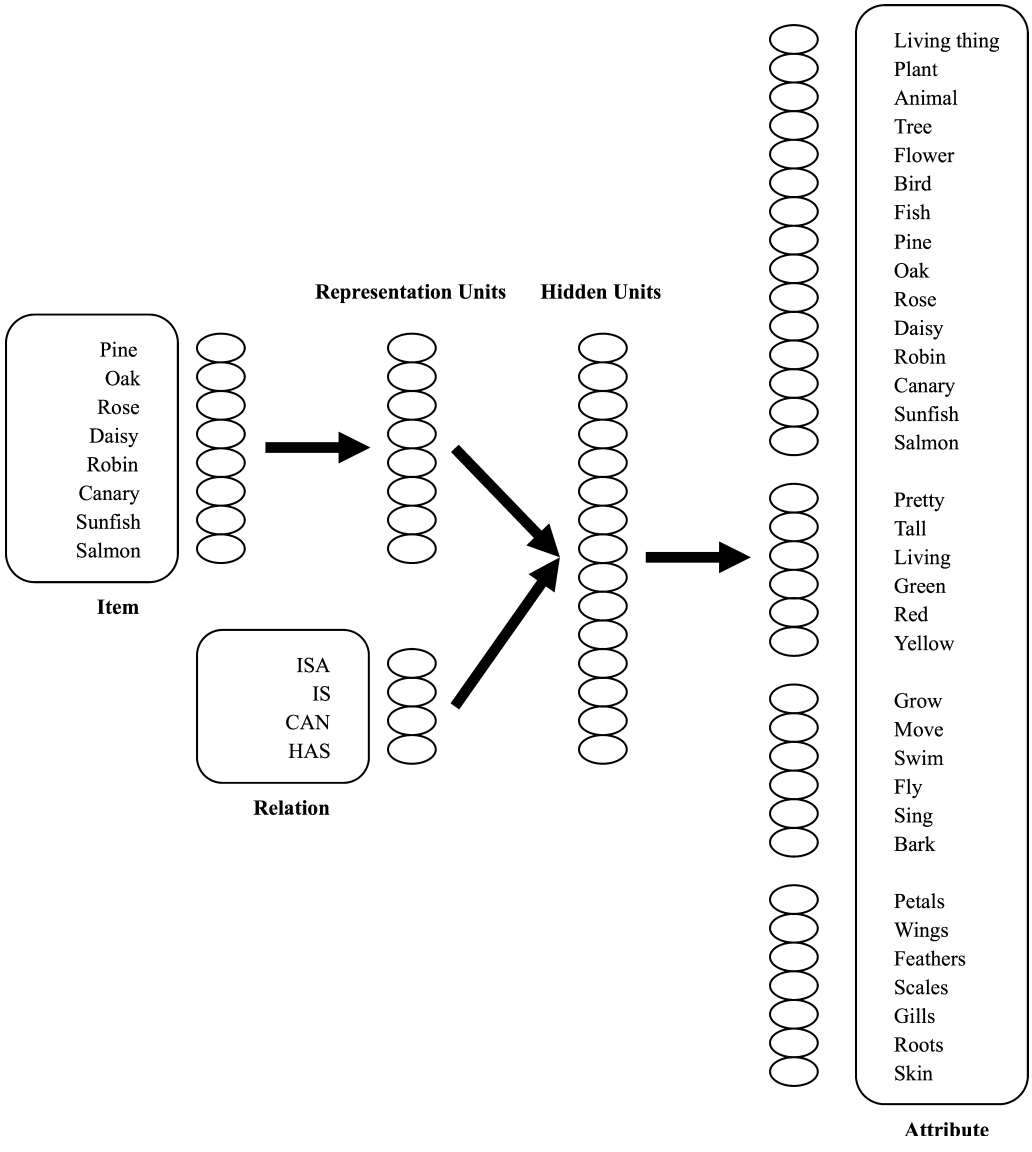

Fig. 3. Feedforward semantic network. Input to the network consists of an item-relation pair: activation of a unit in the Item layer is set to one, and activation of a unit in the Relation layer is set to one. Inputs to the network pass through the Representation and Hidden Units layers before activating output units. The network is trained to activate all units in the Attribute layer that satisfy the item-relation pair. Solid arrows indicate that all units in the sending layer (to the left) connect to all units in the receiving layer (to the right). Figure adapted from Rogers and McClelland (2004).

The third form of redundancy is also architectural. In earlier versions of this model, all item units project to all representation units. Consequently, inputs activate distributed representations. Alternatively, each item unit could project to one unique representation unit. In this case, localist representations would be maintained across the representation layer. We varied connectivity between the item and representation layers to test whether the architectural redundancy afforded by distributed representations did, in fact, enhance the robustness of the network against simulated damage.

8.2. Simulation and results

We implemented semantic networks based on Rogers and McClelland (2004). In versions with distributed representations, all item units connected with all representation units. Connection weights were initialized to random values between −0.9 and 0.9, and were allowed to change during training. The network contained six, eight, or ten units in the representation layer. In the version with localist representations, each item unit connected with one unique representation unit. Connection weights were fixed at 1.0. The network contained eight units in the representation layer. All networks were trained through 5,000 epochs. In each epoch, the network was presented with 32 examples formed by crossing the eight items and four relations. The learning rate was set to 0.1. Results are based on 100 simulations of each network.

The simulation involved a single perturbation: number of damaged units in the representation layer. After training, we simulated neural damage by setting representation units' incoming activation from the item layer to zero. We varied the number of damaged units in the representation layer from zero to six, and tested the ability of the networks to activate positive attributes of each concept. Some attributes apply to all concepts in the network (e.g., *is living*), others to superordinate categories (e.g., *can move*), others to intermediate categories (e.g., *has wings*), and still others to individual concepts (e.g., *can sing*). Responses were counted as correct if activation of the target output unit exceeded 0.5. Test accuracy provided the values of S and F for the robustness quantification (Eq. 2). To simplify matters, we set tolerance for incorrect responses to 1.0 (T in Eq. 2).

During training, versions of the network with distributed representations correctly learned general attributes first, followed by superordinate attributes, intermediate attributes, and finally individual attributes. The time-course of acquisition of the different attributes in the network with eight hidden units was virtually identical to Rogers and McClelland (2004). Following training, all versions of the network correctly activated all attributes. In other words, in the absence of perturbation, representational and architectural redundancy had no effect on eventual functionality.

Having validated the re-implemented model, we applied the perturbation. As the number of damaged representation units increased (Fig. 4), the functionality of all networks and for each type of attribute (excluding universal attributes) dropped. Because representation units convey item inputs to subsequent network layers, functionality should logically fall with the number of damaged units. The rate of change in function-

ality was modulated by the degrees of representational and architectural redundancy in the network. The network with the most representation units maintained highest functionality across the range of perturbation. Of the networks that contained eight representation units, the network with distributed representations maintained greater functionality than did the network with localist representations.

We tested whether representational redundancy enhanced the robustness of the semantic network against damage. As seen in Fig. 4, universal attributes were most robust against damage, followed by superordinate attributes, intermediate attributes, and individual attributes (Table 3). Likewise, universal attributes were most stable, and individual attributes were least stable (Table 3).

These results arise in part from the amount of activation spreading from relation units to attributes. The network acquires large positive weights from relation units to universal attributes via the hidden layer. Activation of a relation unit (e.g., *isa*) is sufficient to activate the universal attribute (e.g., *living thing*). The network acquires moderate positive weights from relation units to superordinate attributes via the hidden layer. Activation of a relation unit (e.g., *isa*) moderately activates superordinate attributes (e.g., *plant* and *animal*), but activation of an item unit is necessary to disambiguate the correct response. Lastly, the network acquires small positive weights from relation units to intermediate and individual attributes via the hidden layer. Activation of a relation unit only weakly activates intermediate and individual attributes, and activation of an item unit is necessary to drive the response.

In networks with distributed representations, these results also arise from the similarity structure among items across the representation layer (Rogers & McClelland, 2004). An individual item (e.g., *canary*) occupies a narrow region of the representation space, its intermediate category (e.g., *bird*) occupies a broader region, and its superordinate category (e.g., *animal*) occupies the broadest region. Consequently, when an item's representation is degraded, the representation is still likely to correctly activate superordinate attributes, it is less likely to activate intermediate attributes, and it is least likely to activate individual attributes.

Next, we tested whether the number of representation units in networks with distributed representations affected robustness. A 4 (attribute) × 3 (number of representation units) repeated measures ANOVA revealed main effects of attribute, $F(3, 891) = 7{,}279$, $p < .0001$, $\eta^2 = 0.96$, and of number of representation units, $F(2, 297) = 1{,}134$, $p < .0001$, $\eta^2 = 0.88$. Networks with more representation units were most robust against damage. Representational redundancy interacted with this form of architectural redundancy, $F(6, 891) = 413$, $p < .0001$, $\eta^2 = 0.74$. More fragile sources of knowledge (i.e., superordinate, intermediate, and individual attributes) were especially vulnerable to damage in networks with few representation units. Findings pertaining to stability mirrored the results of the robustness analysis (Table 3).

These results held when controlling for the proportion of damaged units. Considering only the case where half of the representation units were damaged (3/6 in the network with six representation units, 4/8 in the network with eight representation units,

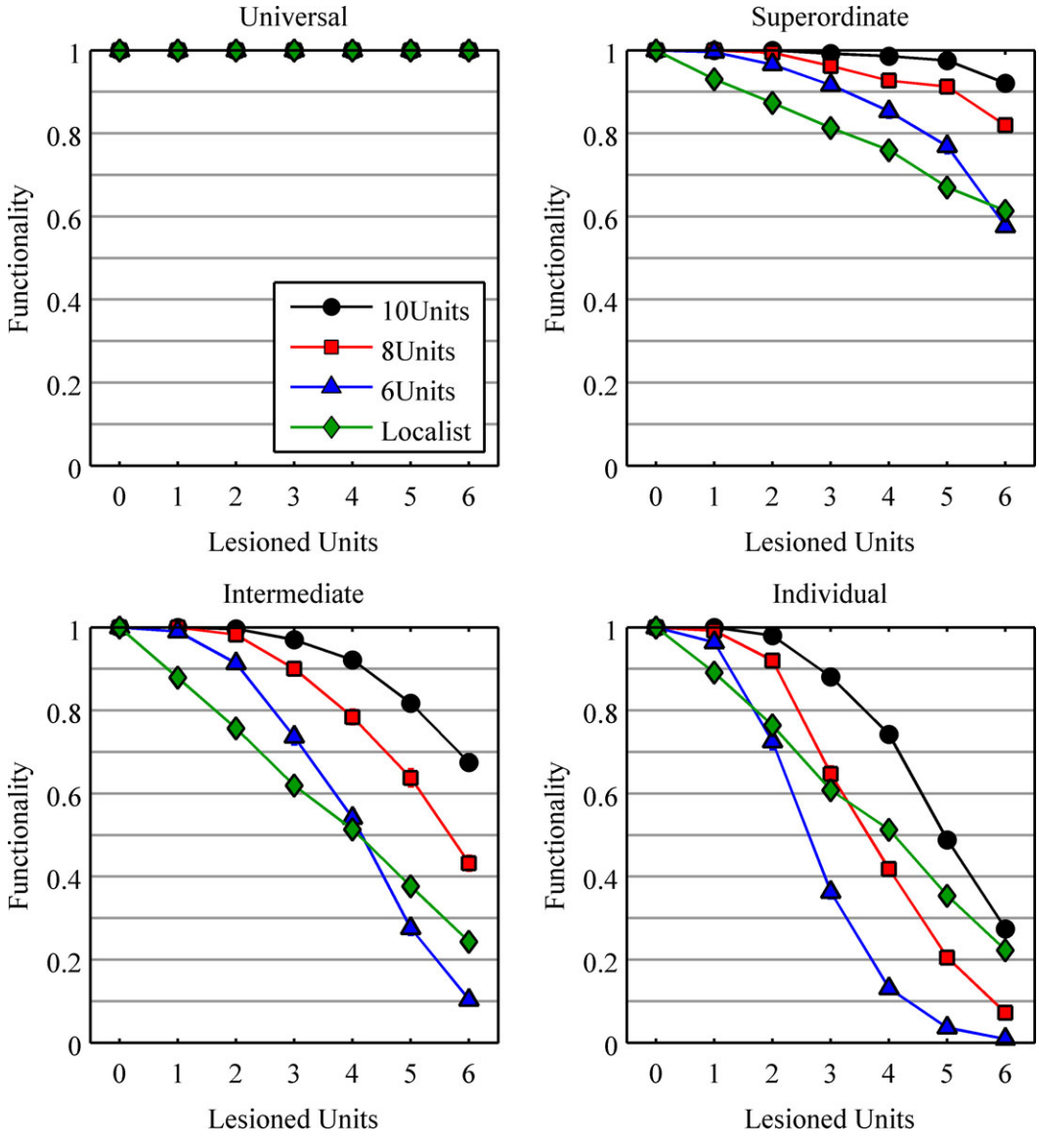

Fig. 4. Semantic memory simulations. Functional of the semantic network as number of lesioned representation units increases from 0 to 6. Panels show attributes that apply to categories at different levels of generality: universal (living thing), superordinate (plants or animals), intermediate (tree, flower, bird, or fish), and individual (pine, oak, rose, daisy, robin, canary, sunfish, or salmon). Lines within each panel correspond to networks with between 6 and 10 representation units, and network with localist representation of items.

and 5/10 in the network with 10 representation units), a 4 (attribute) × 3 (number of representation units) ANOVA revealed a main effect of number of representation units, $F(2, 297) = 18$, $p < .0001$, $\eta^2 = .11$. Across superordinate, intermediate, and individual

Table 3
Robustness and stability of semantic network models based on the generality of attribute and the number of units within the network's representation layer

	Attribute Level							
	Universal		Superordinate		Intermediate		Individual	
No. of Units	Robustness	Stability	Robustness	Stability	Robustness	Stability	Robustness	Stability
10	1.00	1.00	0.98	0.92	0.91	0.71	0.77	0.40
8	1.00	1.00	0.95	0.82	0.82	0.51	0.61	0.20
6	1.00	1.00	0.87	0.63	0.65	0.25	0.46	0.11
Localist (8)	1.00	1.00	0.82	0.70	0.62	0.44	0.63	0.41

attributes, the network with 10 representation units performed about 10% better than did the network with six representation units. The size of the effect increased to 20% when we compared networks with 12 and four representation units.

Lastly, we tested whether the type of representation, distributed or localist, affected the robustness of networks with eight units in the representation layer. A 4 (attribute) × 2 (representation) repeated measures ANOVA revealed main effects of attribute, $F(3, 594) = 4984$, $p < .0001$, $\eta^2 = 0.96$, and representation, $F(1, 198) = 650$, $p < .0001$, $\eta^2 = 0.77$. The network with distributed representations was more robust against damage. Architectural redundancy interacted with this form of representational redundancy, $F(3, 594) = 436$, $p < .0001$, $\eta^2 = 0.69$. Superordinate knowledge was very robust against damage in both networks, and individual knowledge was not especially robust in either network. But superordinate and intermediate sources of knowledge were far more robust in the network with distributed representations. Findings pertaining to stability mirrored the results of the robustness analysis (Table 3).

Performance of the network with localist representations depends on two factors. First is the amount of activation spreading from relation units to attributes. As in the networks with distributed representations, relation units acquire strong positive weights with universal attributes via the hidden layer, moderate positive weights with superordinate attributes, and weak positive weights with intermediate and individual attributes. In the absence of input from the item layer, the network always activates universal attributes above threshold, it activates superordinate attributes above threshold 50% of the time, and it never activates intermediate or individual attributes above threshold. The second factor is the probability that the active unit in the representation layer is damaged. As the number of damaged units increases, so too does this likelihood. Performance for intermediate and individual attributes tracks the combinatorial probability of choosing the critical unit from a pool of m units given n draws. When the critical unit is not damaged, performance is unaffected. But when the critical unit is damaged, performance drops to zero. This contrasts with the effects of damage in versions of the network with distributed representations, where each additional unit lost causes a slight dip in performance.

8.3. Discussion

The number of neocortical neurons decreases by about 10% during adulthood (Pakkenberg & Gundersen, 1997). How can the brain achieve reliable computations with such unreliable hardware? One solution to this problem is to allocate extra neurons, or units, to internal representations. We found that semantic networks with fewer than six representation units learned to properly activate attributes for a collection of items. But these networks were very sensitive to simulated neural damage. The performance of networks with a greater number of representation units, though equivalent in the absence of perturbation, was far more robust and stable against simulated damage. This conclusion held even after controlling for the proportion of damaged units. In other words, as the number of representation units increases, each becomes *disproportionately* less important. When representations are distributed, each unit can participate in multiple representations. Increasing the number of units increases the degree of redundancy in the network, thereby offsetting the effects of damage.

Another way to achieving reliable computations is to use the same number of units but in different ways. We contrasted networks where each item unit projected to one representation unit or to multiple representation units. The former pattern of connectivity maintains localist representations of items across the representation layer, but the latter re-encodes localist representations in a distributed fashion. Although distributed representations degrade gracefully (Hinton & Sejnowski, 1986), it was unclear whether versions of the network with this type of representation would actually be more robust. The constant accumulation of small performance decrements in the networks with distributed representations could exceed the large, but low probability decrements caused by damaging individual units in the network with localist representations. This was not the case; the network with distributed representations was more robust and stable against damage. These simulations confirm that *redundancy*, inherent in the architecture of the semantic network, enhances the robustness of high-level representations against the loss of low-level units.

Localist representations do not preclude redundancy. Multiple nodes may independently represent an entity (Bowers, 2009; Page, 2000). For example, Konorsky (1967) suggested that the number of redundant units dedicated to a stimulus is proportional to its importance. This highlights a trade space; increasing redundancy enhances robustness but necessitates more processing units. A related tradeoff in information theory is between the efficiency of data transmission (i.e., source coding), and the reliability of data transmission (i.e., channel coding). If each unit participates in multiple representations, as with distributed representations, reliability can be increased without greatly decreasing efficiency (Hinton, McClelland, & Rumelhart, 1986; O'Reilly, 1996).[5]

Aside from these forms of architectural redundancy, these simulations feature a type of representational redundancy in terms of the generality of attributes. Universal attributes were most robust and stable, and individual attributes were least robust and stable. The loss of knowledge in semantic dementia occurs in the reverse order that it was acquired: from specific to general concepts (Warrington, 1975). Networks with distributed represen-

tations produced varying degrees of attribute robustness like those actually seen in patients with semantic dementia.

9. Simulation 3: Adaptability in strategy selection

The domain of mathematical cognition contains abundant examples of strategic variability. For example, when solving addition problems, children sometimes start from zero and count both numbers. Other times, they start from the larger number and count up. Still other times, they retrieve the answer from memory. This variability is not an artifact of one child consistently using one strategy and another child using a different strategy. Within a session, a single child may use as many as five different addition strategies (Siegler, 1987). This variability is also not unique to children performing mental arithmetic. People of all ages exhibit strategic variability in mathematical and non-mathematical tasks alike (for a review, see Siegler et al., 1996).

To the extent that problem features favor different strategies, variability can enhance robustness. To realize the advantages conferred by having a collection of specialized strategies, however, the individual needs an adaptive choice mechanism. One solution to this strategy selection problem is reinforcement learning (Erev & Barron, 2005; Rieskamp & Otto, 2006; Sutton & Barto, 1998; Walsh & Anderson, 2013). After the individual enacts a strategy and receives feedback, a *reward prediction error* is computed. The prediction error equals the difference between the reward the individual actually received and the reward they expected. Prediction errors are used to update the estimated value of the selected strategy. When expectations are revised in this way, the individual can learn to accurately predict the utility of each strategy. This form of *adaptability* enhances the robustness of the cognitive system against variations in problem features.

9.1. Numerosity judgment

Luwel, Verschaffel, Onghena, and de Corte (2003, 2005), studied strategy selection in a numerosity judgment task. Participants viewed a 7×7 grid that contained filled and empty squares. In each trial, they were asked how many squares were filled. Participants spontaneously adopted two strategies. The first, *addition* strategy, involves counting each filled square and stating the cumulative sum. The second, *subtraction* strategy, involves counting each empty square and subtracting the cumulative sum from the total number of squares.

In their experiments, Luwel, Verschaffel, Onghena, and de Corte (2003), Luwel, Lemaire, and Verschaffel (2005) used the choice/no-choice method (Siegler & Lemaire, 1997). In no-choice trials, participants were told which strategy to use. No-choice trials provide information about strategy performance characteristics that is not biased by problem selection artifacts. In choice trials, participants were allowed to select a strategy. Choice trials provide information about the adaptiveness of participants' selections. Results from no-choice trials showed that as the number of filled squares increased, solution times for the addition strategy rose and solution times for the sub-

traction strategy fell. In agreement with these outcomes, participants favored the addition strategy during choice trials when few squares were filled, and they favored the subtraction strategy when many squares were filled.

9.2. Task model

We built mathematical models of the addition and subtraction strategies. Both models contained an intercept term to capture the duration of cognitive events that were constant across trials, and both contained a slope term to capture the duration of cognitive events associated with counting each additional square. On the basis of data from no-choice trials (Luwel et al., 2005), we estimated separate intercept terms for the addition (0.15 s) and subtraction strategies (2.56 s), and one slope term (0.42 s/square) for both strategies.[6] That the intercept for the subtraction strategy was greater is sensible given that this strategy includes the extra step of subtracting the cumulative number of empty squares from the total number of squares.

We paired these strategies with an adaptive selection mechanism that used reinforcement learning. The mechanism calculates the expected utility of each strategy as a linear function of the current input,

$$U_a = \sum_{j=1:n} w_{j,a} \cdot I_j \tag{8}$$

where $w_{j,a}$ is the learned weight between input j and strategy a, and I_j is the activation of input j. There are two network inputs: a tonic unit with activation set to one, and an experiment unit that codes the number of filled squares in a continuously varying fashion from 0.0 (none filled) to 1.0 (all filled).[7] At the start of each trial, the model calculates the expected utility of the two strategies based on the number of filled squares, and selects the strategy with greater utility.[8]

After responding, the network receives reward that is exponentially discounted according to the duration of the trial ($\lambda < 1.0$),

$$\text{reward} = 1 \cdot \lambda^{\text{duration}} \tag{9}$$

In the studies by Luwel et al. (2003, 2005), participants did not receive feedback about whether their responses were correct. When outcomes are equivalent, as in this case, discounting causes individuals to favor choices that reduce delays. A large literature documents discounting effects in humans and animals (for a review, see Frederick, Loewenstein, & O'Donoghue, 2002).

Prediction error (δ) was calculated as the difference between the actual reward received (Eq. 9) and the expected reward for using the chosen strategy (Eq. 8). Prediction error was used to adjust network weights. The learning rate (α) controlled the step size of updates,[9]

$$w_{j,a} \leftarrow w_{j,a} + \alpha \cdot \delta \cdot I_j \tag{10}$$

To summarize, the adaptive selection mechanism learned to associate inputs with each strategy's utility, which was a decreasing function of its duration.

9.3. Simulation and results

We tested four models: two contained a single strategy (*addition* or *subtraction*), and two contained both strategies along with a *random* or an *adaptive* selection mechanism. The simulations involved a single perturbation: number of filled squares. For each model, we simulated performance of 500 participants on numerosity judgment problems where the number of filled squares ranged from 1 to 49.

Response times for the addition model increased with the number of filled squares, and response times for the subtraction model increased with the number of empty squares (Fig. 5). Neither model responded quickly over the full range of problems. Response times for the random model, though minimally affected by the number of filled squares, were uniformly slow. Finally, response times for the adaptive model hugged the minimum duration of the addition and subtraction strategies. Of the four models, only the adaptive model responded quickly to all problems. This was reflected in its lower mean reaction time across all problems (Adaptive: 7.23 s; Random: 11.63 s; Addition: 10.71 s; Subtraction: 12.70 s).

The adaptive model seemed most robust. To confirm this analytically, we performed a robustness analysis of the four models. Because there were no incorrect responses, we focused on response times. This emphasis reflects the assumption that there are pressures on cognitive systems to act quickly and correctly (Wickelgren, 1977). Depending on the task and context, speed and accuracy contribute differentially to functionality. We binned responses over one-second intervals ranging from 1 to 20 seconds, and penalized all responses that took longer than 1 second,

$$\text{Functionality} = \left(R_1 - \left[\frac{R_2}{T_2} + \frac{R_3}{T_3} + \ldots \frac{R_{20}}{T_{20}} \right] \right) \tag{11}$$

We set tolerances for responses occurring during each interval as a decreasing, linear function of the interval's duration, $T_n = 1.00 - 0.05 \cdot (\text{duration} - 1)$. Effectively, tolerance was least for the slowest responses. Following normalization (Eq. 3), we computed robustness and stability over the range of problems.

As the number of filled squares increased, functionality of the addition model dropped sharply. Conversely, as the number of empty squares increased, functionality of the subtraction model dropped sharply. Functionality of the random model was low when few squares or many squares were filled (i.e., when either strategy performed poorly), and functionality of the adaptive model remained high except for when an intermediate number of squares were filled (i.e., when neither strategy performed well). The regions over which the addition and subtraction strategies were most effective did not overlap. Nor did the regions over which they were least effective. The adaptive

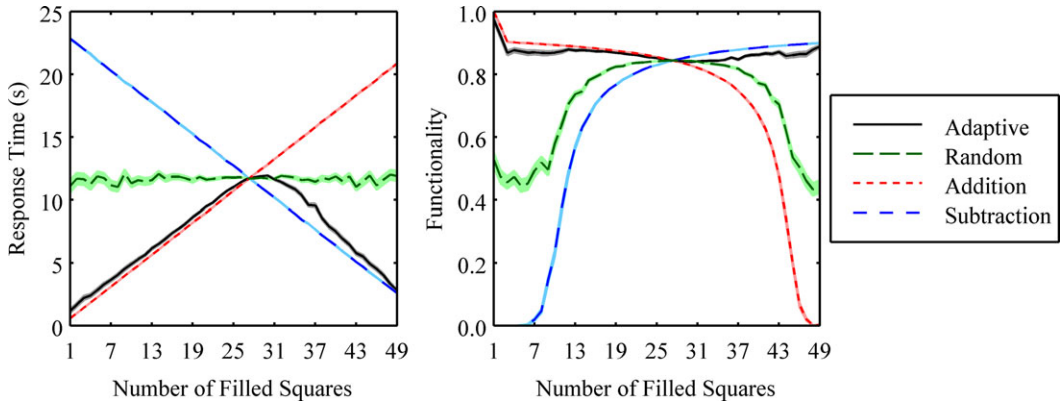

Fig. 5. Numerosity judgment simulations. Average solution times (left) and functionality (right) of adaptive model, random model, addition model, and subtraction model for problems with different numbers of filled squares. Shaded regions show ±1 standard error of the mean.

selection mechanism learned when to deploy each strategy, yielding the most robust and stable performance of the four models (Table 4). Though we only tested the models for grids with 49 squares based on Luwel et al. (2003, 2005), the relative robustness and stability advantages of the adaptive model hold over grids of all sizes.

The robustness analysis yielded three other outcomes. First, the subtraction strategy was less robust than the addition strategy. This is because the intercept term was greater for the subtraction strategy than for the addition strategy (2.56 s vs. 0.15 s). Second, the subtraction strategy was less stable than the addition strategy even though variability in their solution times across problems was identical—the same slope term (0.42 s/square) was used to predict the additional time to count each empty or filled square. The mapping between reaction times (or accuracy) and functionality can be non-linear. In this example, functionality drops more steeply as response times increase (Eq. 11). Because the subtraction strategy has a larger intercept term, equivalent slowing yields lower stability than for the addition strategy. Third, the random model has low stability even though its mean functionality changes little with the number of filled squares. Stability measures two sources of variability: variability across scenarios, and variability within scenarios. For the addition and subtraction models, low stability was mainly driven by variability across scenarios (e.g., the time to apply *addition* varies greatly between problems that contain 1 or 49 filled squares). For the random model, low stability was mainly driven by variability within scenarios (e.g., the time to respond to a problem with five filled squares varies greatly between the *addition* and *subtraction* strategies, yet both strategies are selected with equal frequency). For the adaptive model, variability within scenarios was low because the model consistently selected one strategy, and variability across scenarios was low because the model consistently selected the best strategy.

Table 4
Mean robustness and stability of numerosity judgment models, with standard deviations in parentheses

Model	Robustness	Stability
Adaptive	0.865 (0.037)	0.865 (0.124)
Random	0.701 (0.038)	0.392 (0.065)
Addition	0.743 (0.010)	0.462 (0.033)
Subtraction	0.653 (0.011)	0.321 (0.024)

9.4. Discussion

Strategic variability is a pervasive characteristic of performance in domains such as arithmetic (Siegler, 1987), spelling (Rittle-Johnson & Siegler, 1999), and judgment and decision making (Gigerenzer & Gaissmaier, 2011). When multiple strategies exist, people favor those that are best aligned with specific problem features. The choice mechanism we described, an adaptive network that learns from experience, can be extended to each of these domains. These simulations demonstrate how *adaptability* enhances the robustness of behavior against variations in problem features.

The adaptive choice mechanism in this example is based on reinforcement learning, a technique developed in computer science (Sutton & Barto, 1998). Behavioral studies furnished early support for reinforcement learning in the form of the "law of effect" (Thorndike, 1911). Single-cell recordings provided further support by showing that the firing rate of dopamine neurons depends on the difference between actual and expected outcomes (Schultz, 1998). Neuroimaging experiments have since extended this result to humans by establishing that blood–oxygen level-dependent responses also mirror reward prediction errors (O'Doherty, 2004), as do electrical signals originating from neural regions implicated in behavioral selection (Walsh & Anderson, 2012). Thus, reinforcement learning is a principled and general way to achieve adaptability.

This is not the only way to achieve adaptability, however. Distinction is drawn between model-free and model-based RL (Walsh & Anderson, 2014). In model-free RL, the approach used here, the individual learns action values directly. In model-based RL, the individual learns which outcomes different actions lead to, and the values of those outcomes. Model-free RL relies on reward prediction errors to learn action values, whereas model-based RL relies on associative learning mechanisms to acquire state transition and reward probabilities. There are normative arguments for including both types of control in cognitive agents, there is evidence that the brain implements each (Daw et al., 2005), and there has been recent progress toward an integrated computational account of both (Veksler, Myers, & Gluck, 2014). Ohlsson (2011) proposed nine different modes of learning. Each of these modes, along with RL and associative learning, supports adaptability in different circumstances. Of these, model-free RL has proven especially successful at accounting for strategy selection effects across multiple trials, as in the numerosity judgment task studied here (Erev & Barron, 2005; Rieskamp & Otto, 2006; Sutton & Barto, 1998; Walsh & Anderson, 2013).

Participants' behavior in the choice condition of Luwel et al. (2003, 2005) closely resembled that of the adaptive model. As the number of filled squares increased, reaction times rose and then fell. Participants' overt verbal reports also suggested that they favored the addition strategy on problems with few filled squares, and that they favored the subtraction strategy on problems with many filled squares. The process measure used in this experiment, verbal protocols, was instrumental in establishing that participants possessed a mixture-of-strategies, and that they switched strategies when the number of filled squares exceeded a critical value. The adaptive model also learns when to switch among strategies. In this way, it best captures the degree of robustness that participants exhibit in the numerosity judgment task.

The adaptive model embodies a degree of redundancy: either strategy can be used to solve any problem, however slowly. But in this example, redundancy alone did not increase robustness. The random model also contained multiple strategies, yet it performed no better than the addition or subtraction models. Adaptability requires redundancy, but it also requires a way to sense the state of the world, and a method for acquiring and representing knowledge about which strategies to use in different states.

10. General discussion

Environmental variability drives the requirement for robustness in biology and engineering, but not exclusively there. Cognitive systems also exist within variable environments. The emerging view from psychology is that robustness is a key property of cognition as well, motivating the question of how cognitive systems achieve a degree of robustness. We found that the same general mechanisms that enhance robustness in biological and engineered systems—system control, redundancy, and adaptability—foster robust cognition. Our simulations provided concrete demonstrations that these mechanisms safeguard cognitive capacities against a range of perturbations. In psychomotor vigilance, system control reduces the deleterious effects of fatigue. In semantic memory, architectural redundancy protects high-level representations from damage to low-level units. Lastly, in numerosity judgment, adaptability preserves performance despite variations in problem features.

10.1. Other possible mechanisms

Our list of mechanisms for robust cognition is not exhaustive. Others likely exist. For example, in biology, robustness has been associated with purging and anti-redundancy, neutrality and sloppiness, conflict management, error detection and repair, and modularity (Flack et al., 2012). To focus on but one of these, the encapsulation of functions into separable units increases robustness by restricting damage to local parts of the organism (Hartwell, Hopfield, Leibler, & Murray, 1999). Such *modularity* is evident across multiple levels of biological organization (Csete & Doyle, 2002).

Likewise, engineers include isolated segments in structural designs to arrest collapse at segment borders (Canisius, 2011; Starossek & Haberland, 2012).

The concept of modularity has been influential in neuroscience. Measures of neural connectivity indicate that the brain consists of a hierarchy of modules (Bassett & Gazzaniga, 2011; Park & Friston, 2013). Each module contains densely intraconnected nodes, which share functional specialization and anatomical location.

The concept of modularity has also influenced the study of cognition. Fodor's (1983) ideas on the modularity of mind are clearly relevant; however, the theorizing of Pylyshyn (1999) on cognitive impenetrability and of Simon (1996) on the nature of complex systems like the mind as nearly decomposable, hierarchic systems are also foundational in discussions regarding cognitive modularity. Modularity is evident in contemporary integrated cognitive architectures (Anderson, 2007), and Kurzban (2010) explores the role of evolutionary selection pressures in producing modular minds. Although we have not formally quantified robustness for varying degrees of modularity in this study, there seems to be accumulating evidence in these prior publications that modularity enhances the robustness of cognitive systems, as it does for biological and engineered systems.

10.2. The cost of robustness

We have argued that robustness enhances fitness. Yet robustness is costly (Gluck et al., 2012). For example, a function that depends on multiple redundant components may be less likely to fail than a function that depends on one component. But the reduction in the probability of failure comes at the cost of increased resource demands. The brain ameliorates these costs through pluripotency, a type of structure–function relationship in which one region can participate in multiple functions (Just & Varma, 2007; Noppeney et al., 2004). The pluripotentiality of the neural system permits redundancy without duplication. Likewise, internal representations of concepts are encoded by the activation of multiple units, each of which participates in multiple representations. Thus, distributed representations are robust because they do not depend on a single unit, and they are efficient because each unit participates in multiple representations (Hinton & Sejnowski, 1986; O'Reilly, 1998; Rumelhart & Todd, 1993).

Resource demands are not the only cost of robustness. The study of complex systems has given rise to the idea that robustness is a conserved quantity (Csete & Doyle, 2002; Kitano, 2007). In order for robustness to increase in one area, it must decrease elsewhere. For example, modern airplanes are robust against anticipated perturbations such as weather variation, but fragile against unanticipated perturbations such as power failure. In other words, a system cannot be categorically or universally robust. Rather, a system is robust against some sources of variation, but fragile against others.

The fragility of the cognitive system is evident in addiction. The neurotransmitter dopamine has been implicated in reward learning (Schultz, 1998). The reinforcing effects of dopamine underlie adaptability. Certain drugs of addiction hijack the dopamine response, however, producing behavior that is extremely maladaptive *and* extremely

robust against intervention (Redish, Jensen, & Johnson, 2008). Although we have focused on robustness in this study, an understanding of cognitive vulnerabilities is equally important. Such an understanding can guide the development of interventions and technologies to further augment cognitive robustness.

10.3. Advancing a quantification of robustness

Throughout the field of psychological science, neural circuits, cognitive capacities, and decisions heuristics are often said to be robust. Yet robustness is never directly measured. One contribution of this study is a quantification of robustness (see also Walsh et al., 2013). This quantification serves as a starting point, but it leaves some issues unresolved.

First, statements about robustness must be made with respect to specific perturbations. How should the types and ranges of perturbations be selected? In naturalistic settings, perturbations and their likelihood (i.e., probability (x) in Eq. 4) may be determined from observation. In applied settings, anticipated perturbations may be specified in the design process. In both cases, the choice of perturbation must be clearly articulated and defensible. This specification challenge is not unique to our quantification. For example, normalized maximum likelihood estimation, a tool for model evaluation, requires selecting and restricting the ranges of potentially unbounded dependent variables (Myung, Navarro, & Pitt, 2006). In our quantification, the related challenge is to restrict the ranges of independent variables.

Second, should functionality be treated as a continuous variable, and robustness as the integral of functionality across scenarios? An alternate approach is to compute the proportion of scenarios in which performance remains within allowable risk, and is thus viable (Hafner, Koeppl, Hasler, & Wagner, 2009; Larhlimi, Blachon, Selbig, & Nikoloski, 2011). Yet another, more conservative approach is to treat robustness as functionality in the worst-case scenario (i.e., minimax; Bitmore, 2009). Identifying the strengths and weaknesses of these approaches, and when to use each, awaits further analysis.

Third and finally, does stability matter? In our simulations, stability correlated with robustness. We failed to dissociate robustness and stability because maximum functionality of the candidate models was always similar. As a consequence, systems that exhibited the smallest loss of functionality were most robust, and necessarily most stable. This need not be the case. For example, a statistical rule based on optimal weighting of cues may perform very well when applied to items from the training set but generalize poorly. Alternatively, a heuristic based on the value of a single cue may perform moderately well when applied to items from the training set, and also generalize well (Gigerenzer & Gaissmaier, 2011). Though the two may be equally robust, the heuristic would be preferred for its stability. In fact, predictability of performance, a construct related to stability, is a key determinant in people's perception of trustworthiness and reliance on automation (Lee & See, 2004).

10.4. Universal laws and invariant properties

Chater and Brown (2008) argue for universal laws of cognition, laws that cut across cognitive domains. Such laws reflect convergent adaptations to the structure of the environment (see also, Anderson, 1991). Our view of robustness as a property of cognition resonates with this position. We argue that the reoccurring collection of mechanisms for robustness reflect evolutionary adaptations to the structure of the environment. Because cognitive systems exist in a variable world, the need for robustness is a critical constraint in a computational theory of cognition (Marr, 1982).

Simon (1996) drew a distinction between the natural sciences and the artificial sciences. In this dichotomy, the artificial sciences are concerned with phenomena that are contingent on the interaction between a system's purposes or goals and the environment in which it exists. In contrast, the natural sciences do not involve the study of goal-driven systems. Simon's research was motivated predominantly by the proposal that bounded rationality (1982, 1996) is the necessity, the invariant, that emerges from the contingencies of artificial systems. Perhaps robustness is the invariant property that emerges from selection pressures operating on all complex systems, whether natural or artificial. The merit of this proposal must be evaluated by others, but the idea serves as a foundation for our interest in domain-general mechanisms that produce robust cognition.

10.5. Future directions

To advance the study of robustness in cognitive science, we advocate three practices. First, quantification of robustness should supplant vague qualitative claims. The word *robust* appears throughout the psychological literature (Walsh et al., 2013). Unfortunately, it is used inconsistently and ambiguously. This impedes scientific and technical progress in measuring and achieving robustness. Quantification enhances the transparency of theory predictions, and provides an objective way to express and compare the performance of different systems.

Our quantification is not meant to replace metrics of model fit (e.g., response times, response accuracy, etc.). In fact, our quantification can be applied equally well to simulated data from models and also to empirical data from experiment participants. In this way, robustness can be treated as a dependent measure to assess the correspondence between predicted and observed performance. As noted throughout the simulations, certain model variants exhibited a degree of robustness akin to experiment participants, and others did not. To perform this type of comparison, however, one must first conduct the necessary studies with humans.

This brings us to our second general recommendation. Studying robustness requires adopting different experiment paradigms. To show that a cognitive system is robust against a source of variation, one must evaluate the system in multiple states; for example, testing a participant who is rested and then fatigued. The mechanisms we identified protect performance against perturbation, and so cannot be demonstrated with a single

condition. This is evident in our first and second simulations: In the absence of perturbation, system control did not affect functionality in psychomotor vigilance, and redundancy did not affect functionality in semantic memory. In addition, to show that a mechanism enhances robustness against a source of variation, one must disrupt the mechanism and then apply the perturbation. If the mechanism enhances robustness, performance will be impaired in the perturbed state when the mechanism is disrupted.

Third and finally, although cognitive models have successfully captured the behavior of individuals in isolated scenarios, they often prove fragile when exposed to new tasks or objectives (Anderson & Lebiere, 2003; Gluck & Pew, 2005; Newell, 1990). Although robustness may be a universal law or an invariant property of cognitive systems, it is not typically a property of computational cognitive models. Incorporating the mechanisms we identified here into cognitive architectures and the knowledge available to them will enhance the robustness of cognitive models.

Some progress has already been made toward this objective. Different integrated cognitive architectures incorporate these mechanisms to varying extents. For example, in the ACT-R model of the PVT (Gunzelmann et al., 2009), the cognitive system decreases the utility threshold for action selection upon detecting fatigue. Other architectures monitor physiological and emotional states to select high-level goals that regulate those states (Bach, 2009; Gratch & Marsella, 2007). These forms of system control protect performance against fatigue and other potentially threatening cognitive moderators. In addition, the 4CAPS architecture consists of cortical areas that become active during task performance (Just & Varma, 2007). Each area can perform multiple cognitive functions, and each cognitive function can be performed by multiple areas. This redundancy allows 4CAPS to maintain cognitive capacities even when cortical areas responsible for them are damaged or unavailable. Lastly, the Soar cognitive architecture was recently expanded to include reinforcement learning (Nason & Laird, 2005). This allows Soar agents to acquire statistical information about the past success of their actions to facilitate greater adaptability in operator selection. System control, redundancy, and adaptability are certainly relevant for cognitive architectures, but that is not the only message of this study. Inclusion of these mechanisms in cognitive models more generally is a necessary step toward human-level robustness.

Acknowledgments

This research was performed while the first author held a National Research Council Research Associateship Award with the Air Force Research Laboratory's Cognitive Models and Agents Branch. We thank AFOSR for grant 13RH06COR to the second author. The authors thank AFRL for supporting this research and thank members of the Robust Decision Making Strategic Technology Team and the Cognitive Models and Agents Branch for stimulating conversation and critical thinking on this topic. Glenn Gunzelmann, Chris Myers, Jorge Zuniga, and three anonymous reviewers provided constructive comments on an earlier version of this manuscript.

Notes

1. Though we focus on human cognition in this study, our position is that robustness is a key property of non-human cognitive systems as well.
2. This list is not exhaustive, but these are the most studied mechanisms of robustness (for others, see Flack et al., 2012).
3. FT is a notational variation in the threshold adjustment mechanism previously introduced in Gunzelmann et al. (2009).
4. The standard approach to evaluating cognitive models is to compute goodness of fit metrics using dependent variables such as response time and accuracy. By these metrics, the fit of the model to the PVT data is exceptional (Gunzelmann et al., 2009; Walsh et al., *submitted*). An alternate approach, suggested here, is to compute the robustness of human behavior, and to compare it with model robustness.
5. Another relevant tradeoff is between adequately capturing outputs and overfitting the data. An excessively flexible model might respond correctly to training stimuli, but generalize poorly. As the number of units in the representation layer increases, so does this potential for overfitting (Geman, Bienenstock, & Doursat, 1992). Other factors such as weight limits and training procedure also affect the tendency for overfitting.
6. These models closely matched solution time data from adults in the no-choice condition of Luwel et al. (2005), $r^2 = .99$, MSE = 0.009. To produce variability in response durations, we treated the time to count each square as a uniform random variable with a range equal to its mean. We also treated the intercept term as a uniform random variable with a range equal to its mean.
7. The tonic unit determines the baseline activation of each strategy's output unit in the absence of external input.
8. Logistically distributed noise with a mean of zero and a standard deviation of 0.05 was added to utility values to produce choice variability.
9. We set the discounting parameter (λ) to 0.9 and the learning rate (α) to 0.3. We obtained similar results even after reducing the discounting parameter and learning rate by a factor of two. Before testing the adaptive network, we trained it on one pass through the problem set.

References

Agrawal, A. A., Laforsch, C., & Tollrian, R. (1999). Transgenerational induction of defenses in animals and plants. *Nature*, *401*, 60–63.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–484.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe*. New York: Oxford University Press.

Anderson, J. R., & Lebiere, C. (2003). The Newell Test for a theory of cognition. *Behavioral and Brain Sciences*, *26*, 587–637.

Bach, J. (2009). *Principles of synthetic intelligence: PSI: An architecture of motivated cognition*. New York: Oxford University Press.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*, 48–69.

Barkai, N., & Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, *387*, 913–917.

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, *15*, 200–209.

Berg, J., Tymoczko, J., & Stryer, L. (2002). *Biochemistry* (5th ed.). New York: W. H. Freeman.

Bitmore, K. (2009). *Rational decisions*. Princeton, NJ: Princeton University Press.

Bontempi, F., Giuliani, L., & Gkoumas, K. (2007). Handling the exceptions: Dependability of systems and structural robustness, 3rd International Conference on Structural Engineering, Mechanics and Computation (pp. 104–110). Cape Town, South Africa..

Borbély, A. A., & Achermann, P. (1999). Sleep homeostasis and models of sleep regulation. *Journal of Biological Rhythms*, *6*, 559–570.

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences, USA*, *110*, 1628–1633.

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

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*, 220–251.

Brown, L. E., & Rosenbaum, D. A. (2002). Motor control: Models. In L. Nadel (Ed.), *Encyclopedia of cognitive science, 3* (pp. 127–133). London: Macmillan.

Cain, N. H., Barreiro, A. K., Shalden, M., & Shea-Brown, E. (2013). Neural integrators for decision making: A favorable tradeoff between robustness and sensitivity. *Journal of Neurophysiology*, *109*, 2542–2559.

Canisius, T. D. G. (Ed.) (2011). Structural robustness design for practising engineers. COST Action TU0601. Available at http://www.cost.eu/media/publications/11-38-Robustness-of-Structures-Final-Report-of-COST-Action-TU0601. Accessed March 14, 2014.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*, 36–67.

Chee, M. W., & Choo, W. C. (2004). Functional imaging of working memory after 24 hr of total sleep deprivation. *Journal of Neuroscience*, *24*, 4560–4567.

Chee, M. W. I., Tan, J. C., Zheng, H., Parimal, S., Weissman, D. H., Zagorodnov, V., & Dinges, D. F. (2008). Lapsing during sleep deprivation is associated with distributed changes in brain activation. *Journal of Neuroscience*, *28*, 5519–5528.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, *8*, 240–247.

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

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. London: Academic Press.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences*, *4*, 423–431.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, *17*, 652–655.

Doran, S. M., van Dongen, H. P., & Dinges, D. F. (2001). Sustained attention performance during sleep deprivation: Evidence of state instability. *Archives of Italian Biology: Neuroscience*, *139*, 253–267.

Drummond, S. P., Brown, G. G., Salamat, J. S., & Gillin, J. C. (2004). Increasing task difficulty facilitates the cerebral compensatory response to total sleep deprivation. *Sleep*, *27*, 445–451.

Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, *25*, 117–129.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences, USA*, *98*, 13763–13768.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*, 912–931.

Flack, J. C., Hammerstein, P., & Krakauer, D. C. (2012). Robustness in biological and social systems. In J. R. Stevens & P. Hammerstein (Eds.), *Evolution and the mechanisms of decision making. Strüngmann Forum Report*, vol. 11, J. Lupp, series ed. (pp. 129–149). Cambridge, MA: MIT Press.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Friston, K. J., & Price, C. J. (2003). Degeneracy and redundancy in cognitive anatomy. *Trends in Cognitive Sciences*, *7*, 151–152.

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

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gluck, K. A., McNamara, J. M., Brighton, H., Dayan, P., Kareev, Y., Krause, J., Kurzban, R., Selten, R., Stevens, J. R., Voelkl, B., & Wimsatt, W. C. (2012). Robustness in a variable environment. In J. R. Stevens & P. Hammerstein (Eds.), *Evolution and the mechanisms of decision making. Strüngmann Forum Report*, vol. 11, J. Lupp, series ed. (pp. 195–214). Cambridge, MA: MIT Press.

Gluck, M. A., & Myers, C. E. (1997). Psychobiological models of hippocampal function in learning and memory. *Annual Review of Psychology*, *48*, 481–514.

Gluck, K. A., & Pew, R. W. (Eds.) (2005). *Modeling human behavior with integrated cognitive architectures: Comparison, evaluation, and validation*. Mahwah, NJ: Erlbaum.

Goodwin, G. C., Graebe, S. F., & Salgado, M. E. (2001). *Control system design*. Englewood Cliffs, NJ: Prentice Hall.

Gratch, J., & Marsella, S. (2007). The architectural role of emotion in cognitive systems. In W. D. Gray (Ed.), *Integrated models of cognitive systems* (pp. 230–242). New York: Oxford University Press.

Gray, R. (2011). Looming auditory collision warnings for driving. *Human Factors*, *53*, 63–74.

Gunzelmann, G., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep deprivation and sustained attention performance. Integrating mathematical and cognitive modeling. *Cognitive Science*, *33*, 880–910.

Habeck, C., Rakitin, B. C., Moeller, J., Scarmeas, N., Zarahn, E., Brown, T., & Stern, Y. (2004). An event-related fMRI study of the neurobehavioral impact of sleep deprivation on performance of a delayed-match- to-sample task. *Cognitive Brain Research*, *18*, 306–321.

Hafner, M., Koeppl, H., Hasler, M., & Wagner, A. (2009). "Glocal" robustness analysis and model discrimination for circadian oscillators. *PLOS Computational Biology*, *5*, 1–10.

Hagen, E. H., & Hammerstein, P. (2005). Evolutionary biology and the strategic view of ontogeny: Genetic strategies in the life course. *Research in Human Development*, *2*(87), 101.

Hammerstein, P., Hagen, E. H., Herz, A. V. M., & Herzel, H. (2006). Robustness: A key to evolutionary design. *Biological Theory*, *1*, 90–93.

Hammerstein, P., & Stevens, J. R. (2012). Six reasons for invoking evolution in decision theory. In P. Hammerstein & J. R. Stevens (Eds.), *Evolution and the mechanisms of decision making. Strüngmann Forum Report*, vol. 11, J. Lupp, series ed. (pp. 1–17). Cambridge, MA: MIT Press.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, *402*, C47–C52.

Hatakeyama, T. S., & Kaneko, K. (2012). Generic temperature compensation of biological clocks by autonomous regulation of catalyst concentration. *Proceedings of the National Academy of Sciences, USA*, *109*, 8109–8114.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Huaco, D. R., Bowders, J. J., & Loehr, J. E. (2012). Method to develop target levels of reliability for design using LRFD. In Transportation Research Board 91st Annual Meeting (No. 12-4327). Washington, DC.

Ioannou, P. A., & Sun, J. (1996). *Robust adaptive control*. Upper Saddle River, NJ: Prentice-Hall.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Jordan, M. I. (1996). Computational aspects of motor control and motor learning. In J. Moody, S. Hanson, & R. Lippmann (Eds.), *Advances in neural information processing* (pp. 985–993). San Mateo, CA: Morgan Kaufmann.

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience*, *7*, 153–191.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*, 718–727.

Kitano, H. (2004). Biological robustness. *Nature Neuroscience*, *5*, 826–837.

Kitano, H. (2007). Toward a theory of biological robustness. *Molecular Systems Biology*, *3*, 1–7.

Kohonen, T. (1984). *Self-organization and associative memory*. New York: Springer-Verlag.

Konorsky, J. (1967). *Integrative activity of the brain: An interdisciplinary approach*. Chicago: University of Chicago Press.

Kurzban, R. (2010). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton, NJ: Princeton University Press.

Larhlimi, A., Blachon, S., Selbig, J., & Nikoloski, Z. (2011). Robustness of metabolic networks: A review of existing definitions. *Biosystems*, *106*, 1–8.

Lashley, K. S. (1950). In search of the engram. In J. F. Danielli, & R. Brown (Eds.), *Physiological mechanisms in animal behavior* (pp. 454–482). New York: Academic Press.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 50–80.

Lim, J., & Dinges, D. F. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, *1129*, 305–322.

Luwel, K., Lemaire, P., & Verschaffel, L. (2005). Children's strategies in numerosity judgment. *Cognitive Development*, *20*, 448–471.

Luwel, K., Verschaffel, L., Onghena, P., & de Corte, E. (2003). Analysing the adaptiveness of strategy choices using the choice/no-choice method: The case of numerosity judgment. *European Journal of Cognitive Psychology*, *15*, 511–537.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*, 6–21.

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

Marr, D. (1971). Simple memory: A theory for archicortex. *Proceedings of the Royal Society of London, Series B*, *841*, 23–81.

Marr, D. (1982). *Vision*. New York: W. H. Freeman.

McCauley, P., Kalachev, L. V., Mollicone, D. J., Banks, S., Dinges, D. F., & van Dongen, H. P. A. (2013). Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep*, *36*, 1987–1997.

Mummery, C. J., Patterson, K., Wise, R. J. S., Vandenbergh, R., Price, C. J., & Hodges, J. R. (1999). Disrupted temporal lobe connections in semantic dementia. *Brain*, *122*, 61–73.

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

Nason, S., & Laird, J. E. (2005). Soar-RL: Integrating reinforcement learning with soar. *Cognitive Systems Research*, *6*, 51–59.

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

NISTIR 7396 (2007) *Best practices for reducing the potential for progressive collapse in buildings*. Gaithersburg, MD: National Institute of Standards and Technology (NIST).

Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*, *205*, 433–442.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*, 769–776.

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. New York: Cambirdge University Press.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*, 455–462.

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

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, *23*, 443–512.

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

Park, H. J., & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science*, *342*, 1238411.

Paus, T., Zatorre, R. J., Hofle, N., Caramanos, Z., Gotman, J., Petrides, M., & Evans, A. C. (1997). Time-related changes in neural systems underlying attention and arousal during the performance of an auditory vigilance task. *Journal of Cognitive Neuroscience*, *9*, 392–408.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552.

Portas, C. M., Rees, G., Howseman, A. M., Josephs, O., Turner, R., & Frith, C. D. (1998). A specific role for the thalamus in mediating the interaction of attention and arousal in humans. *Journal of Neuroscience*, *18*, 8979–8989.

Pratt, R. (2000). *Flight control systems: Practical issues in design and implementation*. United Kingdom: Institute of Engineering and Technology.

Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences*, *6*, 416–421.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–423.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences*, *31*, 415–487.

Reinman, G., Ayer, T., Davan, T., Devore, M., Finley, S., Glanovski, J., Gray, L., Hall, B., Jones, C., Learned, A., Mesaros, E., Morris, R., Pinero, S., Russo, R., Stearns, E., Teicholz, M., Teslik-Welz, W., & Yudichak, D. (2012). Design for variation. *Quality Engineering*, *24*, 317–345.

Reppert, S. M., & Weaver, D. R. (2002). Coordination of circadian timing in mammals. *Nature*, *418*, 935–941.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

Rittle-Johnson, B., & Siegler, R. S. (1999). Learning to spell: Variability, choice, and change in children's strategy use. *Child Development*, *70*, 332–348.

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

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*, 362–380.

Saper, C. B., Scammell, T. E., & Lu, J. (2005). Hypothalamic regulation of sleep and circadian rhythms. *Nature*, *437*, 1257–1263.

Sastry, S., & Bodson, M. (1989). *Adaptive control: Stability, convergence, and robustness*. Englewood Cliffs, NJ: Prentice-Hall.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Scott, J. J., & Gray, R. (2008). A comparison of tactile, visual, and auditory warnings for rear-end collision prevention in simulated driving. *Human Factors*, *50*, 264–275.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*, 250–264.

Siegler, R. S., Adolph, K. E., & Lemaire, P. (1996). Strategy choices across the life span. In L. R. Reder (Ed.), *Implicit memory and metacognition* (pp. 79–121). Mahwah, NJ: Erlbaum.

Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*, 71–92.

Simon, H. A. (1982). Comments. In M. S. Clark & S. T. Fiske (Eds.), *Affect and cognition: The 17th Annual Carnegie Symposium on cognition* (pp. 333–342). Mahwah, NJ: Erlbaum.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Squire, L. (1987). *Memory and brain*. New York: Oxford University Press.

Starossek, U., & Haberland, M. (2012). Robustness of structures. *International Journal of Lifecycle Performance Engineering*, *1*, 3–21.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*, 548–565.

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

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*, 907–915.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2014). SAwSu: An integrated model of associative and reinforcement learning (online early access). *Cognitive Science*, *38*, 580–598.

Wagner, A. (2005). *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.

Walsh, M. M., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology*, *58*, 416–440.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*, 1870–1884.

Walsh, M. M., & Anderson, J. R. (2013). Electrophysiological responses to feedback during the application of abstract rules. *Journal of Cognitive Neuroscience*, *25*, 1986–2002.

Walsh, M. M., & Anderson, J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological Bulletin*, *140*, 466–486.

Walsh, M. M., Einstein, E. H., & Gluck, K. A. (2013). A quantification of robustness. *Journal of Applied Research in Memory and Cognition*, *2*, 137–148.

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. A. (2014). Comparing accounts of psychomotor vigilance impairment due to sleep loss. 36th Annual Conference of the Cognitive Science Society. Quebec City, Canada.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*, 433–441.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*, 635–657.

Weinstein, M. C., & Fineberg, H. V. (1980). *Clinical decision analysis*. Philadelphia, PA: W. B. Saunders.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, *1*, 209–216.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.

Wyatte, D., Curran, T., & O'Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience*, *24*, 2248–2261.

Yi, T. M., Huang, Y., Simon, M. I., & Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences, USA*, *97*, 4649–4653.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*, 513–523.