Behavioral and Brain Sciences

cambridge.org/bbs

Target Article

Cite this article: Lieder F, Griffiths TL. (2020) Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* **43**, e1: 1–60. doi:10.1017/S0140525X1900061X

Target Article Accepted: 18 January 2019
Target Article Manuscript Online: 4 February 2019
Commentaries Accepted: 22 May 2019

Keywords:
bounded rationality; cognitive biases; cognitive mechanisms; cognitive modeling; representations; resource rationality

CAMBRIDGE
UNIVERSITY PRESS

Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources

Falk Lieder[a] and Thomas L. Griffiths[b]

[a]Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany and [b]Departments of Psychology and Computer Science, Princeton University, Princeton, New Jersey 08544, USA
falk.lieder@tuebingen.mpg.de https://re.is.mpg.de
tomg@princeton.edu https://psych.princeton.edu/person/tom-griffiths

Abstract

Modeling human cognition is challenging because there are infinitely many mechanisms that can generate any given observation. Some researchers address this by constraining the hypothesis space through assumptions about what the human mind can and cannot do, while others constrain it through principles of rationality and adaptation. Recent work in economics, psychology, neuroscience, and linguistics has begun to integrate both approaches by augmenting rational models with cognitive constraints, incorporating rational principles into cognitive architectures, and applying optimality principles to understanding neural representations. We identify the rational use of limited resources as a unifying principle underlying these diverse approaches, expressing it in a new cognitive modeling paradigm called *resource-rational analysis*. The integration of rational principles with realistic cognitive constraints makes resource-rational analysis a promising framework for reverse-engineering cognitive mechanisms and representations. It has already shed new light on the debate about human rationality and can be leveraged to revisit classic questions of cognitive psychology within a principled computational framework. We demonstrate that resource-rational models can reconcile the mind's most impressive cognitive skills with people's ostensive irrationality. Resource-rational analysis also provides a new way to connect psychological theory more deeply with artificial intelligence, economics, neuroscience, and linguistics.

1. Introduction

Cognitive modeling plays an increasingly important role in our endeavor to understand the human mind. Building models of people's cognitive strategies and representations is useful for at least three reasons. First, testing our understanding of psychological phenomena by recreating them in computer simulations forces precision and helps to identify gaps in explanations. Second, computational modeling permits the transfer of insights about human intelligence to the creation of artificial intelligence (AI) and vice versa. Third, cognitive modeling of empirical phenomena is a way to infer the underlying psychological mechanisms, which is critical to predicting human behavior in novel situations.

Unfortunately, inferring cognitive mechanisms and representations from limited experimental data is an ill-posed problem, because any behavior could be generated by infinitely many candidate mechanisms (Anderson 1978). Thus, cognitive scientists must have strong inductive biases to infer cognitive mechanisms from limited data. Theoretical frameworks, such as evolutionary psychology (Buss 1995), embodied cognition (Wilson 2002), production systems (e.g., Anderson 1996), dynamical systems theory (Beer 2000), connectionism (Rumelhart & McClelland 1987), Bayesian models of cognition (Griffiths et al. 2010), ecological rationality (Todd & Gigerenzer 2012), and the free-energy principle (Friston 2010) to name just a few, provide researchers guidance in the search for plausible hypotheses. Here, we focus on a particular subset of theoretical frameworks that emphasize developing computational models of cognition: cognitive architectures (Langley et al. 2009), connectionism (Rumelhart & McClelland 1987), computational neuroscience (Dayan & Abbott 2001), and rational analysis (Anderson 1990). These frameworks provide complementary functional or architectural constraints on modeling human cognition. Cognitive architectures, such as ACT-R (Anderson et al. 2004), connectionism, and computational neuroscience constrain the modeler's hypothesis space based on previous findings about the nature, capacities, and limits of the mind's cognitive architecture. These frameworks scaffold explanations of psychological phenomena with assumptions about what the mind can and cannot do. But the space of cognitively feasible mechanisms is so vast that most phenomena can be explained in many different ways – even within the confines of a cognitive architecture.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

As psychologists, we are trying to understand a system far more intelligent than anything we have ever created ourselves; it is possible that the ingenious design and sophistication of the mind's cognitive mechanisms are beyond our creative imagination. To address this challenge, rational models of cognition draw inspiration from the best examples of intelligent systems in computer science and statistics. Perhaps the most influential framework for developing rational models of cognition is rational analysis (Anderson 1990). In contrast to traditional cognitive psychology, rational analysis capitalizes on the *functional* constraints imposed by goals and the structures of the environment rather than the structural constraints imposed by cognitive architectures. Its inductive bias toward rational explanations is often rooted in the assumption that evolution and learning have optimally adapted the human mind to the structure of its environment (Anderson 1990). This assumption is supported by empirical findings that under naturalistic conditions people achieve near-optimal performance in perception (Knill & Pouget 2004; Knill & Richards 1996; Körding & Wolpert 2004), statistical learning (Fiser et al. 2010), and motor control (Todorov 2004; Wolpert & Ghahramani 2000), as well as inductive learning and reasoning (Griffiths & Tenenbaum 2006; 2009). Valid rational modeling provides solid theoretical justifications and enables researchers to translate assumptions about people's goals and the structure of the environment into substantive, detailed, and often surprisingly accurate predictions about human behavior under a wide range of circumstances.

That said, the inductive bias of rational theories can be insufficient to identify the correct explanation and sometimes points modelers in the wrong direction. Canonical rational theories of human behavior have several fundamental problems. First, human judgment and decision-making systematically violate the axioms of rational modeling frameworks such as expected utility theory (Kahneman & Tversky 1979), logic (Wason 1968), and probability theory (Tversky & Kahneman 1973; 1974). Furthermore, standard rational models define optimal behavior without specifying the underlying cognitive and neural mechanisms that psychologists and neuroscientists seek to understand. Rational models of cognition are expressed at what Marr (1982) termed the "computational level," identifying the abstract computational problems that human minds must solve and their ideal solutions. In contrast, psychological theories have traditionally been expressed at Marr's "algorithmic level," focusing on representations and the algorithms by which they are transformed.

Falk Lieder leads the Max Planck Research Group for Rationality Enhancement at the MPI for Intelligent Systems in Tübingen. His current research builds on the theory of resource-rationality to develop a scientific foundation and practical tools for improving the human mind by promoting and supporting cognitive growth, goal setting, and goal achievement.

Thomas L. Griffiths is the Henry R. Luce Professor of Information Technology, Consciousness, and Culture at Princeton University. His research explores connections between psychology and computer science, using ideas from machine learning and artificial intelligence to understand how people solve the challenging computational problems they encounter in everyday life.

This suggests that relying either cognitive architectures or rationality alone might be insufficient to uncover the cognitive mechanisms that give rise to human intelligence. The strengths and weaknesses of these two approaches are complementary – each offers exactly what the other is missing. The inductive constraints of modeling human cognition in terms of cognitive architectures were, at least to some extent, built from the ground up by studying and measuring the mind's elementary operations. In contrast, the inductive constraints of rational modeling are derived from top-down considerations of the requirements of intelligent action. We believe that the architectural constraints of bottom-up approaches to cognitive modeling should be integrated with the functional constraints of rational analysis.

The integration of (bottom-up) cognitive constraints and (top-down) rational principles is an approach that is starting to be used across several disciplines, and initial results suggest that combining the strengths of these approaches results in more powerful models that can account for a wider range of cognitive phenomena. Economists have developed mathematical models of bounded-rational decision-making to accommodate people's violations of classic notions of rationality (e.g., Dickhaut et al. 2009; Gabaix et al. 2006; Simon 1956; C. A. Sims 2003). Neuroscientists are learning how the brain represents the world as a trade-off between accuracy and metabolic cost (e.g., Levy & Baxter 2002; Niven & Laughlin 2008; Sterling & Laughlin 2015). Linguists are explaining language as a system for efficient communication (e.g., Hawkins 2004; Kemp & Regier 2012; Regier et al. 2007; Zaslavsky et al. 2018; Zipf 1949), and more recently, psychologists have also begun to incorporate cognitive constraints into rational models (e.g., Griffiths et al. 2015).

In this article, we identify the rational use of limited resources as a common theme connecting these developments and providing a unifying framework for explaining the corresponding phenomena. We review recent multidisciplinary progress in integrating rational models with cognitive constraints and outline future directions and opportunities. We start by reviewing the historical role of classic notions of rationality in explaining human behavior and some cognitive biases that have challenged this role. We present our integrative modeling paradigm, *resource rationality*, as a solution to the problems faced by previous approaches, illustrating how its central idea can reconcile rational principles with numerous cognitive biases. We then outline how future work might leverage *resource-rational analysis* to answer classic questions of cognitive psychology, revisit the debate about human rationality, and build bridges from cognitive modeling to computational neuroscience and AI.

2. A brief history of rationality

Notions of rationality have a long history and have been influential across multiple scientific disciplines, including philosophy (Harman 2013; Mill 1882), economics (Friedman & Savage 1948; 1952), psychology (Braine 1978; Chater et al. 2006; Griffiths et al. 2010; Newell et al. 1958; Oaksford & Chater 2007), neuroscience (Knill & Pouget 2004), sociology (Hedström & Stern 2008), linguistics (Frank & Goodman 2012), and political science (Lohmann 2008). Most rational models of the human mind are premised on the classic notion of rationality (Sosis & Bishop 2014), according to which people act to maximize their expected utility, reason based on the laws of logic, and handle uncertainty according to probability theory. For instance, rational actor models (Friedman & Savage 1948; 1952) predict that decision-makers select the action $a^\star$ that maximizes their expected utility (Von Neumann & Morgenstern 1944), that is

$$a^\star = \arg\max_a \int u(o) \cdot p(o|a)\, do, \tag{1}$$

where the utility function u measures how good the outcome o is from the decision-maker's perspective and $p(o|a)$ is the conditional probability of its occurrence if action a is taken.

Psychologists soon began to interpret the classic notions of rationality as hypotheses about human thinking and decision-making (e.g., Edwards 1954; Newell et al. 1958) and other disciplines also adopted rational principles to predict human behavior. The foundation of these models was shaken when a series of experiments suggested that people's judgment and decision-making systematically violate the laws of logic (Wason 1968) probability theory (Tversky & Kahneman 1974), and expected utility theory (Kahneman & Tversky 1979). These systematic deviations are known as *cognitive biases*. The well-known anchoring bias (Tversky & Kahneman 1974), base-rate neglect and the conjunction fallacy (Kahneman & Tversky 1972), people's tendency to systematically overestimate the frequency of extreme events (Lichtenstein et al. 1978), and overconfidence (Moore & Healy 2008) are just a few examples of the dozens of biases that have been reported over the last four decades (Gilovich et al. 2002). In many cases the interpretation of these empirical phenomena as irrational errors has been challenged by subsequent analyses (e.g., Dawes & Mulford 1996; Fawcett et al. 2014; Gigerenzer 2015; Gigerenzer et al. 2012; Hahn & Warren 2009; Hertwig et al. 2005). But as reviewed below, cognitive limitations also appear to play a role in at least some of the reported biases. While some of these biases can be described by models such as prospect theory (Kahneman & Tversky 1979; Tversky & Kahneman 1992) such descriptions do not reveal the underlying causes and mechanisms. According to Tversky and Kahneman (1974), cognitive biases result from people's use of fast but fallible cognitive strategies known as *heuristics*. Unfortunately, the number of heuristics that have been proposed is so high that it is often difficult to predict which heuristic people will use in a novel situation and what the results will be.

The undoing of expected utility theory, logic, and probability theory as principles of human reasoning and decision-making has not only challenged the idealized concept of "man as rational animal" but also taken away mathematically precise, overarching theoretical principles for modeling human behavior and cognition. These principles have been replaced by different concepts of "bounded rationality" according to which cognitive constraints limit people's performance so that classical notions of rationality become unattainable (Simon 1955; Tversky & Kahneman 1974). While research in the tradition of Simon (1955) has developed notions of rationality that take people's limited cognitive resources into account (e.g., Gigerenzer & Selten 2002), research in the tradition of Tversky and Kahneman (1974) has sought to characterize bounded rationality in terms of cognitive biases. In the latter line of work and its applications, the explanatory principle of bounded rationality has often been used rather loosely, that is without precisely specifying the underlying cognitive limitations and exactly how they constrain cognitive performance (Gilovich et al. 2002). As illustrated in Figure 1, infinitely many cognitive mechanisms are consistent with this rather vague use of the term "bounded rationality." This raises questions about which of those mechanisms people use, which of them they should use, and how these two sets of mechanisms are related to each other. Answering these questions requires a more precise theory of bounded rationality.

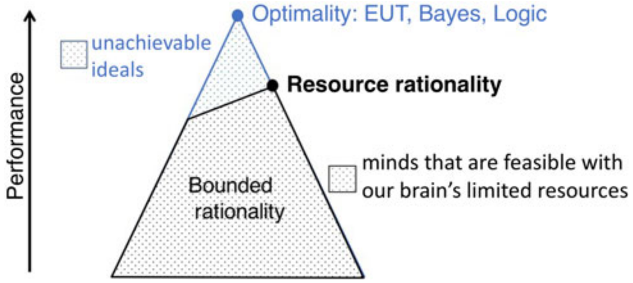

Optimality: EUT, Bayes, Logic
unachievable ideals
Resource rationality
Performance
minds that are feasible with our brain's limited resources
Bounded rationality

Figure 1. Resource rationality and its relationship to optimality and Tversky and Kahneman's concept of bounded rationality. The horizontal dimension corresponds to alternative cognitive mechanisms that achieve the same level of performance. Each dot represents a possible mind. The gray dots are minds with bounded cognitive resources and the blue dots are minds with unlimited computational resources. The thick black line symbolizes the bounds entailed by people's limited cognitive resources. Resource limitations reflect anatomical, physiological, and metabolic constraints on neural information processing as discussed below as time constraints, but they can be modelled at a higher level of abstraction (e.g., in terms of processing speed or multi-tasking capacity). For the purpose of deriving a resource-rational mechanism these constraints are assumed to be fixed. (Some cognitive constraints may change as a consequence of brain development, exhaustion, and many other factors. Sufficiently large changes may warrant the resource-rational analysis to be redone.)

Simon (1955; 1956) famously argued that rational decision strategies must be adapted to both the structure of the environment and the mind's cognitive limitations. He suggested that the pressure for adaptation makes it rational to use a heuristic that selects the first option that is good enough instead of trying to find the ideal option: *satisficing*. Simon's ideas inspired the theory of *ecological rationality*, which maintains that people make adaptive use of simple heuristics that exploit the structure of natural environments (Gigerenzer & Goldstein 1996; Gigerenzer & Selten 2002; Hertwig & Hoffrage 2013; Todd & Brighton 2016; Todd & Gigerenzer 2012). A number of candidate heuristics have been identified over the years (Gigerenzer & Gaissmaier 2011; Gigerenzer & Goldstein 1996; Gigerenzer et al. 1999; Hertwig & Hoffrage 2013; Todd & Gigerenzer 2012) that typically use only a small subset of available information and perform much less computation than would be required to compute expected utilities (Gigerenzer & Gaissmaier 2011; Gigerenzer & Goldstein 1996).

In parallel work, Anderson (1990) developed the idea of understanding human cognition as a rational adaptation to environmental structure and goals pursued within it, creating a cognitive modeling paradigm known as *rational analysis* (Chater & Oaksford 1999) that derives models of human behavior from structural environmental assumptions according to the six steps summarized in Box 1 and Figure 2. Rational process models can be used to connect the computational level of analysis to the algorithmic level of analysis. The principle of resource rationality allows us to derive rational process models from assumptions about a system's function and its cognitive constraints.

Box 1. Rational models developed in this way have provided surprisingly good explanations of cognitive biases by identifying how the environment that people's strategies are adapted to differs from the tasks participants are given in the laboratory and how people's goals often differ from what the experimenter

Box 1 The six steps of rational analysis.

1. Precisely specify what are the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make the minimal assumptions about computational limitations.
4. Derive the optimal behavioral function given items 1 through 3.
5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
6. If the predictions are off, then iterate.

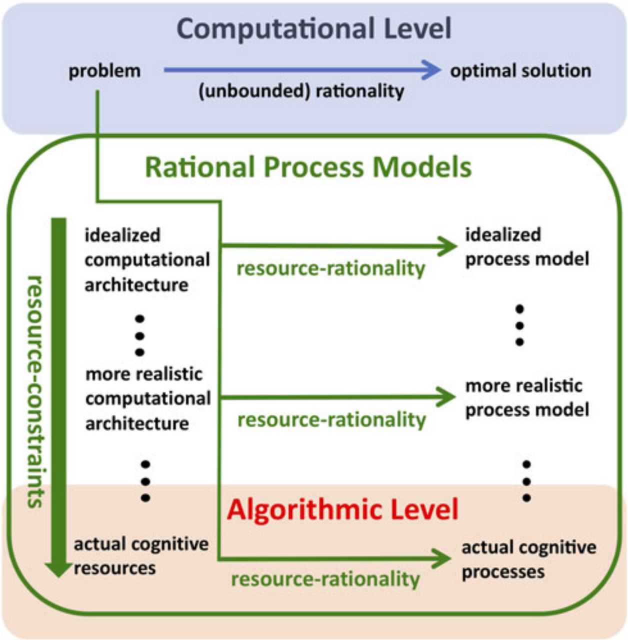


Figure 2. Rational process models can be used to connect the computational level of analysis to the algorithmic level of analysis. The principle of resource rationality allows us to derive rational process models from assumptions about a system's function and its cognitive constraints.

intended them to be; examples include the confirmation bias (Austerweil & Griffiths 2011; Oaksford & Chater 1994), people's apparent misconceptions of randomness (Griffiths & Tenenbaum 2001; Tenenbaum & Griffiths 2001), the gambler's fallacy (Hahn & Warren 2009), and several common logical fallacies in argument construction (Hahn & Oaksford 2007). The theoretical frameworks of ecological rationality and rational analysis are founded on the assumption that evolution has adapted the human mind to the structure of our evolutionary environment (Buss 1995).

Paralleling rational analysis, some evolutionary ecologists seek to explain animals' behavior and cognition as an optimal adaptation to their environments (Houston & McNamara 1999; McNamara & Weissing 2010). This approach predicts the outcome of evolution from optimality principles, but research on how animals forage for food has identified several cognitive biases in their decisions (e.g., Bateson et al. 2002; Latty & Beekman 2010; Shafir et al. 2002). Subsequent work has sought to reconcile these biases with evolutionary fitness maximization by incorporating constraints on animals' information processing capacity and by moving from optimal behavior to optimal decision mechanisms that work well across multiple environments (Dukas 2004; Johnstone et al. 2002).

Research on human cognition faces similar challenges. While it is a central tenet of rational analysis to assume only minimal computational limitations (step 3), the computational constraints imposed by people's limited resources are often substantial (Newell & Simon 1972; Simon 1982) and computing exact solutions to the problems people purportedly solve is often computationally intractable (Van Rooij 2008). For this reason, rational analysis cannot account for cognitive biases resulting from limited resources. A complete theory of bounded rationality must go further in accounting for people's cognitive constraints and limited time.

Fortunately, AI researchers have already developed a theory of rationality that accounts for limited computational resources (Horvitz 1987; Horvitz et al. 1989; Horvitz 1990; Russell 1997; Russell & Subramanian 1995). *Bounded optimality* is a theory for designing optimal programs for agents with performance-limited hardware that must interact with their environments in real time. A program is bounded-optimal for a given architecture if it enables that architecture to perform as well as or better than any other program the architecture could execute instead. This standard is attainable by its very definition. Recently, this idea that bounded rationality can be defined as the solution to a constrained optimization problem has been applied to a particular class of resource-bounded agents: people (Griffiths et al. 2015; Lewis et al. 2014). This leads to a precise theory that uniquely identifies how people should think and decide to make optimal use of their finite time and bounded cognitive resources (see Fig. 1). In the next section, we synthesize and refine these approaches into a paradigm for modeling cognitive mechanisms and representations that we refer to as resource-rational analysis.

3. Resource-rational analysis

While bounded optimality was originally developed as a theoretical foundation for designing intelligent agents, it has been successfully adopted for cognitive modeling (Gershman et al. 2015; Griffiths et al. 2015; Lewis et al. 2014). When combined with reasonable assumptions about human cognitive capacities and limitations, bounded optimality provides a realistic normative standard for cognitive strategies and representations (Griffiths et al. 2015), thereby allowing psychologists to derive realistic models of cognitive mechanisms based on the assumption that the human mind makes rational use of its limited cognitive resources. Variations of this principle are known by various names, including *computational rationality* (Lewis et al. 2014), *algorithmic rationality* (Halpern & Pass 2015), *bounded rational agents* (Vul et al. 2014), *boundedly rational analysis* (Icard 2014), the *rational minimalist program* (Nobandegani 2017), and the idea of rational models with limited processing capacity developed in economics (Caplin & Dean 2015; Fudenberg et al. 2018; Gabaix et al. 2006; C. A. Sims 2003; Woodford 2014) reviewed below. Here, we will refer to this principle as *resource rationality* (Griffiths et al. 2015; Lieder et al. 2012) and advocate its use in a cognitive modeling paradigm called resource-rational analysis (Griffiths et al. 2015).

Figure 1 illustrates that resource rationality identifies the best biologically feasible mind out of the infinite set of bounded-rational minds. To make the notion of resource

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

rationality precise, we apply the principle of bounded optimality to define a resource-rational mind $m^{\star}$ for the brain B interacting with the environment E as

$$m^{\star} = \arg\max_{m \in M_B} \mathbb{E}_{P(T, l_T | E, A_t = m(l_t))}[u(l_T)], \tag{2}$$

where M_B is the set of biologically feasible minds, T is the agent's (unknown) lifetime, its life history $l_t = (S_0, S_1, \cdots, S_t)$ is the sequence of world states the agent has experienced until time t, $A_t = m(l_t)$ is the action that the mind m will choose based on that experience, and the agent's utility function u assigns values to life histories.

Our theory assumes that the cognitive limitations inherent in the biologically feasible minds M_B include a limited set of elementary operations (e.g., counting and memory recall are available but applying Bayes' theorem is not), limited processing speed (each operation takes a certain amount of time), and potentially other constraints, such as limited working memory. Critically, the world state S_t is constantly changing as the mind m deliberates. Thus, performing well requires the bounded optimal mind $m^{\star}$ to not only generate good decisions, but to do so quickly. Since each cognitive operation takes time, bounded optimality often requires computational frugality.

Identifying the resource-rational mind defined by Equation 2 would require optimizing over an entire lifetime, but if we assume that life can be partitioned into a sequence of episodes, we can use this definition to derive the optimal heuristic $h^{\star}$ that a person should use to make a single decision or inference in a particular situation. To achieve this, we decompose the value of having applied a heuristic into the utility of the judgment, decision, or belief update that results from it (i.e., u(result)) and the computational cost of its execution. The latter is critical because the time and cognitive resources expended on any decision or inference (current episode) take away from a person's budget for later ones (future episodes). To capture this, let the random variable cost(t_h, ρ, λ) denote the total opportunity cost of investing the cognitive resources ρ used or blocked by the heuristic h for the duration t_h of its execution, when the agent's cognitive opportunity cost per quantum of cognitive resources and unit time is λ. The resource-rational heuristic $h^{\star}$ for a brain B to use in the belief state b_0 is then

$$\begin{aligned} h^{\star}(s_0, B, E) = \underset{h \in H_B}{\operatorname{argmax}}\, & \mathbb{E}_{P(\text{result}|s_0, h, E, B)}\big[u(\text{result})\big] \\ & - \mathbb{E}_{t_h, \rho, \lambda | h, s_0, B, E}\big[\text{cost}(t_h,\ \rho,\ \lambda)\big], \end{aligned} \tag{3}$$

where H_B is the set of heuristics that brain B can execute and $s_0 = (o,\ b_0)$ comprises observed information about the initial state of the external world (o) and the person's initial belief state b_0. As described below, this formulation makes it possible to develop automatic methods for deriving simple heuristics – like the ones people use – from first principles.

Resource-rational cognitive mechanisms trade off accuracy against effort in an adaptive, nearly optimal manner. This is reminiscent of the proposal that people optimally trade off the time it takes to gather information about prices against its financial benefits (Stigler 1961) but there are two critical differences. The most important difference is that while Stigler (1961) defined a problem to be solved by the decision-maker, Equation 3 defines a problem to be solved by evolution, cognitive development, and life-long learning. That is, we propose that people never have to directly solve the constrained optimization problem defined in Equation 3. Rather, we believe that for most of our decisions the problem of finding a good decision mechanism has already been solved by evolution (Dukas 1998a; McNamara & Weissing 2010), learning, and cognitive development (Siegler & Jenkins 1989; Shrager & Siegler 1998). In many cases the solution $h^{\star}$ may be a simple heuristic. Thus, when people confront a decision they can usually rely on a simple decision rule without having to discover it on the spot. The second critical difference is that while resource rationality is a principle for modeling *internal cognitive mechanisms* (i.e., heuristics) Stigler's information economics defined models of optimal *behavior*. Identifying the optimal behavior (subject to the cost of collecting information) would, in general, require people to perform optimization under constraints in their heads. By contrast, resource-rational analysis will almost invariably favor a simple heuristic over optimization under constraints because it penalizes decision mechanisms by the cost of the mental effort required to execute them and only considers decision-mechanisms that are biologically feasible. That is, while Stigler's information economics focused on the cost of collecting information (e.g., how long it takes to visit different shops to find out how much they charge for a product), resource rationality additionally accounts for the cost of thinking according to one strategy (e.g., evaluating each product's utility in all possible scenarios in which it might be used) versus another (e.g., just comparing the prices).

Equation 3 assumes that all possible outcomes and their probabilities and consequences are known. But the real world is very complex and highly uncertain, and limited experience constrains how well people can be adapted to it. Being equipped with a different heuristic for each and every situation would be prohibitively expensive (Houston & McNamara 1999) – not least because of the difficulty of selecting between them (Milli et al. 2017; 2019). To accommodate these bounds on human rationality, we relax the optimality criterion in Equation 3 from optimality with respect to true environment E to optimality with respect to the information i that has been obtained about the environment through direct experience, indirect experience, and evolutionary adaptation. We can therefore define the boundedly resource-rational heuristic given the limited information i as

$$\begin{aligned} h^{\star}(s_0, B, i) = \operatorname{argmax}_{h \in H_B} \mathbb{E}_{E|i}\Big[& \mathbb{E}_{P(\text{result}|s_0, h, E, B)}\big[u(\text{ result})\big] \\ & - \mathbb{E}_{t_h, \rho, \lambda | h, s_0, B, E}\big[\text{cost}(t_h,\ \rho,\ \ \lambda)\big]\Big]. \end{aligned} \tag{4}$$

Since the mechanisms of adaptation are also bounded, we should not expect people's heuristics to be perfectly resource-rational. Instead, even a resource-rational mind might have to rely on heuristics for choosing heuristics to approximate the prescriptions of Equation 4. Recent work is beginning to illuminate what the mechanisms of strategy selection and adaptation might be (Lieder & Griffiths 2017) but more research is needed to identify how and how closely the mind approximates resource-rational thinking and decision-making.

It is too early to know how resource-rational people really are, but we are optimistic that resource-rational analysis can be a useful methodology for answering interesting questions about cognitive mechanisms – in the same way in which Bayesian modeling is a useful methodology for elucidating what the mind does and why it does what it does (Griffiths et al. 2008; Griffiths et al. 2010). In other words, resource rationality is not a fully fleshed out theory of cognition, designed as a new standard of normativity against which human judgments can be assessed, but a methodological device that allows researchers to translate their assumptions about cognitive constraints and functional requirements into precise mathematical models of cognitive processes and representations.

Resource rationality serves as a unifying theme for many recent models and theories of perception, decision-making, memory, reasoning, attention, and cognitive control that we will review below. While rational analysis makes only minimal assumptions about cognitive constraints, it has been argued that there are many cases where cognitive limitations impose substantial constraints (Simon 1956; 1982). *Resource-rational analysis* (Griffiths et al. 2015) thus extends rational analysis to also consider which cognitive operations are available to people and how costly those operations are in terms of time cognitive resources. This means including the structure and resources of the mind itself in the definition of the environment to which cognitive mechanisms are supposedly adapted. Resource-rational analysis thereby follows Simon's advice that "we must be prepared to accept the possibility that what we call 'the environment' may lie, in part, within the skin of the biological organism" (Simon 1955).

Resource-rational analysis is a five-step process (see Box 2) that leverages the formal theory of bounded optimality introduced above to derive rational process models of cognitive abilities from formal definitions of their function and abstract assumptions about the mind's computational architecture. This function-first approach starts at the computational level of analysis (Marr 1982). When the function of the studied cognitive capacity has been formalized, step 2 of resource-rational analysis is to postulate an abstract computational architecture, that is a set of elementary operations and their costs, with which the mind might realize this function. Next, resource-rational analysis derives the optimal algorithm for solving the problem identified at the computational level with the abstract computational architecture defined in step 2 (Equation 3), thereby pushing the principles of rational analysis toward Marr's algorithmic level (see Fig. 2). The resulting process model can be used to simulate people's responses and reaction times in an experiment. Next, the model's predictions are tested against empirical data. The results can be used to refine the theory's assumptions about the computational architecture and the problem to be solved. The process of resource-rational analysis can then be repeated under these refined assumptions to derive a more accurate process model. Refining the model's assumptions may include moving from an abstract computational architecture to increasingly more realistic models of the mind's cognitive architecture or the brain's biophysical limits. As the assumptions about the computational architecture become increasingly more realistic and the model's predictions become more accurate, the corresponding rational process model should become increasingly more similar to the psychological/neurocomputational mechanisms that generate people's responses (see Fig. 2). The process of resource-rational analysis ends when either the model's predictions are accurate enough or all relevant cognitive constraints have been incorporated sensibly. This process makes resource-rational analysis a methodology for reverse-engineering cognitive mechanisms (Griffiths et al. 2015).

Resource-rational analysis can be seen as an extension of rational-analysis from predicting *behavior* from the structure of the external environment to predicting *cognitive mechanisms* from *internal cognitive resources* and the external environment. These advances allow us to translate our growing understanding of the brain's computational architecture into more realistic models of psychological processes and mental representations. Fundamentally, it provides a tool for replacing the traditional method of developing cognitive process models – in which a theorist imagines ways in which different processes might combine to capture behavior – with a means of automatically deriving hypotheses about cognitive processes from the problem people have to solve and the resources they have available to do so.

Box 2 The five steps of resource-rational analysis. Note that a resource-rational analysis may stop in step 5 even when human performance substantially deviates from the resource-rational predictions as long as reasonable attempts have been made to model the constraints accurately based on the available empirical evidence. Furthermore, refining the assumed computational architecture can also include modeling how the brain might approximate the postulated algorithm.

1. Start with a computational-level (i.e., functional) description of an aspect of cognition formulated as a problem and its solution.
2. Posit which class of algorithms the mind's computational architecture might use to approximately solve this problem, the cost of the computational resources used by these algorithms, and the utility of more accurately approximating the correct solution.
3. Find the algorithm in this class that optimally trades off resources and approximation accuracy (Equation 3 or 4).
4. Evaluate the predictions of the resulting rational process model against empirical data.
5. Refine the computational-level theory (step 1) or assumed computational architecture and its constraints (step 2) to address significant discrepancies, derive a refined resource-rational model, and then reiterate or stop if the model's assumptions are already sufficiently realistic.

Deriving resource-rational models of cognitive mechanisms from assumptions about their function and the cognitive architecture available to realize them (step 2) is the centerpiece of resource-rational analysis (Griffiths et al. 2015). This process often involves manual derivations (e.g., Lieder et al. 2012; 2014), but it is also possible to develop computational methods that discover complex resource-rational cognitive strategies automatically (Callaway et al. 2018a; 2018b; Lieder et al. 2017).

Resource-rational analysis combines the strengths of rational approaches to cognitive modeling with insights from the literature on cognitive biases and capacity limitations. We argue below that this enables resource-rational analysis to leverage mathematically precise unifying principles to develop psychologically realistic process models that explain and predict a wide range of seemingly unrelated cognitive and behavioral phenomena.

4. Modeling capacity limits to explain cognitive biases: case studies in decision-making

In this section, we review research suggesting that the principle of resource rationality can explain many of the biases in decision-making that led to the downfall of expected utility theory. Later, we will argue that the same conclusion also holds for other areas of human cognition. Extant work has augmented rational models with different kinds of cognitive limitations and costs, including costly information acquisition and limited attention, limited representational capacity, neural noise, finite time, and

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

limited computational resources. The following sections review resource-rational analyses of the implications of each of these cognitive limitations in turn, showing that each can account for a number of cognitive biases that expected utility cannot. This brief review illustrates that resource rationality is an integrative framework for connecting theories from economics, psychology, and neuroscience.

4.1 Costly information acquisition and limited attention

People tend to have inconsistent preferences and often fail to choose the best available option even when all of the necessary information is available (Kahneman & Tversky 1979). Previous research has found that many of these violations of expected utility theory might result from the fact that acquiring information is costly (Bogacz et al. 2006; Gabaix et al. 2006; Lieder et al. 2017; Sanjurjo 2017; C. A. Sims 2003; Verrecchia 1982). This cost could include an explicit price that people must pay to purchase information (e.g., Verrecchia 1982), the opportunity cost of the decision-maker's time (e.g., Bogacz et al. 2006; Gabaix et al. 2006) and cognitive resources (Shenhav et al. 2017), the mental effort of paying attention (C. A. Sims 2003), and the cost of overriding one's automatic response tendencies (Kool & Botvinick 2013). Regardless of the source of the cost, we can define resource-rational decision-making as using a mechanism achieving the best possible tradeoff between the expected utility and cost of the resulting decision (see Equation 4).

Rather than trying to evaluate all of their options people tend to select the first alternative they encounter that they consider good enough. For instance, when given the choice between seven different gambles a person striving to win at least $5 may choose the second one without even looking at gambles 3–7 because all of its payoffs range from $5.50 to $9.75. This heuristic is known as *Satisficing* (Simon 1956). Satisficing can be interpreted as the solution to an optimal stopping problem, and Caplin et al. (2011) showed that satisficing with an adaptive aspiration level is a bounded-optimal decision strategy for certain decision problems where information is costly. This analysis can be cast in exactly the form of Equation 3, where the utility of the final outcome trades off against the cost of gathering additional information.

Curiously, people also fail to consider all alternatives even when information can be gathered free of charge. This might be because people's attentional resources are limited. The theory of rational inattention (C. A. Sims 2003; 2006) explains several biases in economic decisions, including the inertia, randomness, and abruptness of people's reactions to new financial information, by postulating that people allocate their limited attention optimally. For instance, the limited attention of consumers may prevent them from becoming more frugal as the balance of their bank account drops, even though that information is freely available to them. Furthermore, the rational inattention model can also explain the seemingly irrational phenomenon that adding an additional alternative can increase the probability that the decision-maker will choose one of the already available options (Matějka & McKay 2015).

The rational inattention model discounts all information equally, but people tend to focus on a small number of relevant variables while neglecting others completely. To capture this, Gabaix (2014) derived which of the thousands of potentially relevant variables a bounded-optimal decision-maker should attend to depending on their variability, their effect on the utilities of alternative choices, and the cost of attention. The resulting *sparse max model* generally attends only to a small subset of the variables, specifies how much attention each of them should receive, replaces unobserved variables by their default values, adjusts the default values of partially attended variables toward their true values, and then chooses the action that is best according to its simplified model of the world. The sparse max model can be interpreted as an instantiation of Equation 4, and Gabaix (2014) and Gabaix et al. (2006) showed that the model's predictions capture how people gather information and predicts their choices better than expected utility theory. In subsequent work, Gabaix extended the sparse max model to sequential decision problems (Gabaix 2016) to provide a unifying explanation for many seemingly unrelated biases and economic phenomena (Gabaix 2017).

People tend to consider only a small number of possible outcomes – often focusing on the worst-case and the best-case scenarios. This can skew their decisions towards irrational risk aversion (e.g., fear of air travel) or irrational risk seeking (e.g., playing the lottery). This may be a consequence of people rationally allocating their limited attention to the most important eventualities (Lieder et al. 2018a).

4.1.1 Noisy evidence and limited time

Noisy information processing is believed to be the root cause of many biases in decision-making (Hilbert 2012). Making good decisions often requires integrating many pieces of weak or noisy evidence over time. However, time is limited and valuable, which creates pressure to decide quickly. The principle of resource rationality has been applied to understand how people trade off speed against accuracy to make the best possible use of their limited time in the face of noisy evidence. Speed-accuracy trade-offs have been most thoroughly explored in perceptual decision-making experiments where people are incentivized to maximize their reward rate (points/second) across a series of self-paced perceptual judgments (e.g., "Are there more dots moving to the right or to the left?"). Such decisions are commonly modelled using variants of the drift-diffusion model (Ratcliff 1978), which has three components: evidence generation, evidence accumulation, and choice. The principle of resource rationality (Equation 3) has been applied to derive optimal mechanisms for generating evidence and deciding when to stop accumulating it.

4.1.2 Deciding when to stop

Research on judgment and decision-making has often concluded that people think too little and decide too quickly, but a quantitative evaluation of human performance in perceptual decision-making against a bounded optimal model suggests the opposite (Holmes & Cohen 2014). Bogacz et al. (2006) showed that the drift-diffusion model achieves the best possible accuracy at a required speed and achieves a required accuracy as quickly as possible. The drift diffusion model sums the difference between the evidence in favor of option A and the evidence in favor of option B over time, stopping evidence accumulation when the strength of the accumulated evidence exceeds a threshold. Bogacz et al. (2006) derived the decision threshold that maximizes the decision-maker's reward rate. Comparing to this optimal speed-accuracy trade-off people gather too much information before committing to a decision (Holmes & Cohen 2014). While Bogacz et al. (2006) focused on perceptual decision-making, subsequent work has derived optimal decision thresholds for value-based choice (Fudenberg et al. 2018; Gabaix & Laibson 2005; Tajima et al. 2016).

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

When repeatedly choosing between two stochastically rewarded actions people (and other animals) usually fail to learn to always choose the option that is more likely to be rewarded; instead, they randomly select each option with a frequency that is roughly equal to the probability that it will be rewarded (Herrnstein 1961). To make sense of this, Vul et al. (2014) derived how many mental simulations a bounded agent should perform for each of its decisions to maximize its reward rate across the entirety of its choices. The optimal number of mental simulations turned out to be very small and depends on the ratio of the time needed to execute an action over the time required to simulate it. Concretely, it is bounded-optimal to decide based on only a single sample, which is equivalent to probability matching, when it takes at most three times as long to execute the action as to simulate it. But when the stakes of the decision increase relative to the agent's opportunity cost, then the optimal number of simulations increases as well. This prediction is qualitatively consistent with studies finding that choice behavior gradually changes from probability matching to maximization as monetary incentives increase (Shanks et al. 2002; Vulkan 2000).

4.1.3 Effortful evidence generation

In everyday life, people often must actively generate the evidence for and against each alternative. Resource-rational models postulating that people optimally tradeoff the quality of their decisions against the cost of evidence generation can accurately capture how much effort decision-makers invest under various circumstances (Dickhaut et al. 2009) and the inversely U-shaped relationship between decision-time and decision-quality (Woodford 2014; 2016).

4.2. Computational complexity and limited computational resources

Many models assume that human decision-making is approximately resource-rational subject to the constraints imposed by unreliable evidence and neural noise (e.g., Howes et al. 2016; Khaw et al. 2017; Stocker et al. 2006). However, Beck et al. (2012) argued that the relatively small levels of neural noise measured neurophysiologically cannot account for the much greater levels of variability and suboptimality in human performance. They propose that instead of making optimal use of noisy representations, the brain uses approximations that entail systematic biases (Beck et al. 2012). From the perspective of bounded optimality, approximations are necessary because the computational complexity of decision-making in the real world far exceeds cognitive capacity (Bossaerts & Murawski 2017; Bossaerts et al. 2018). People cope with this computational complexity through efficient heuristics and habits. In the next section, we argue that resource rationality can provide a unifying explanation for each of these phenomena.

4.3. Resource-rational heuristics

More reasoning and more information do not automatically lead to better decisions. To the contrary, simple heuristics that make clever use of the most important information can outperform complex decision-procedures that use large amounts of data and computation less cleverly (Gigerenzer & Gaissmaier 2011). This highlights that resource rationality critically depends on which information is considered and how it is used.

To solve complex decision problems, people generally take multiple steps in reasoning. Choosing those cognitive operations well is challenging because the benefit of each operation depends on which operations will follow: In principle, choosing the best first cognitive operation requires planning multiple cognitive operations ahead. Gabaix and Laibson (2005) proposed that people simplify this intractable meta-decision-making problem by choosing each cognitive operation according to a myopic cost–benefit analysis that pits the immediate improvement in decision quality expected from each decision operation against its cognitive cost (see Equation 3). Gabaix et al. (2006) found that this model correctly predicted people's suboptimal information search behavior in a simple bandit task and explained how people choose between many alternatives with multiple attributes better than previous models.

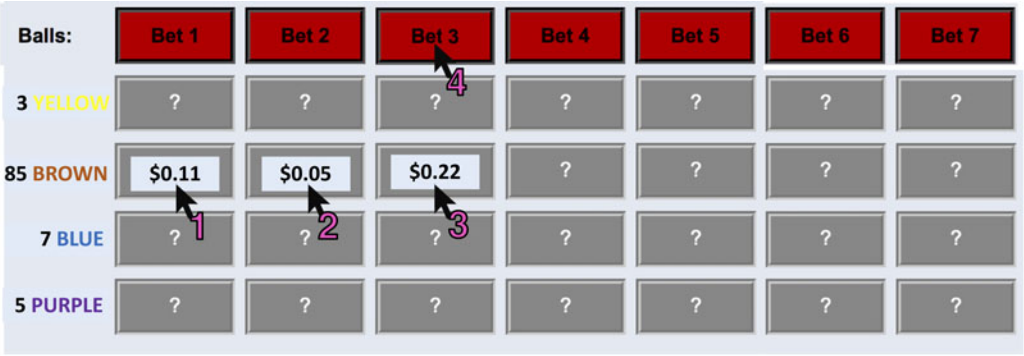
Recent work has developed a non-myopic approach to deriving resource-rational heuristics (Callaway et al. 2018a; Lieder et al. 2017) and previously proposed heuristic models of planning. They also found that people's planning operations achieved about 86% of the best possible trade-off between decision quality and time cost and agreed with the bounded-optimal strategy about 55% of the time. This quantitative analysis offers a more nuanced and presumably more accurate assessment of human rationality than qualitative assessments according to which people are either "rational" or "irrational." Furthermore, this resource-rational analysis correctly predicted how people's planning strategies differ across environments and that their aspiration levels decrease as people gather more information.

This line of work led to a new computational method that can automatically derive resource-rational cognitive strategies from a mathematical model of their function and assumptions about available cognitive resources and their costs. This method is very general and can be applied across different cognitive domains. In an application to multi-alternative risky choice (Lieder et al. 2017), and elucidated the conditions under which they are bounded-optimal. Furthermore, it also led to the discovery of a previously unknown heuristic that combines elements of satisficing and Take-The-Best (SAT-TTB; see Figure 3). A follow-up experiment confirmed that people do use that strategy specifically for the kinds of decision problems for which it is bounded-optimal. These examples illustrate that bounded-optimal mechanisms for complex decision problems generally involve approximations that introduce systematic biases, supporting the view that many cognitive biases could reflect people's rational use of limited cognitive resources.

4.4. Habits

In sharp contrast to the prescription of expected utility theory that actions should be chosen based on their expected consequences, people often act habitually without deliberating about consequences (Dolan & Dayan 2013). The contrast between the enormous computational complexity of expected utility maximization (Bossaerts & Murawski 2017; Bossaerts et al. 2018) and people's limited computational resources and finite time suggests that habits may be necessary for bounded-optimal decision-making. Reusing previously successful action sequences allows people to save substantial amounts of time-consuming and error-prone computation; therefore, the principle of resource rationality in Equation 3 can be applied to determine under which circumstances it is rational to rely on habits.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press



Balls:	Bet 1	Bet 2	Bet 3	Bet 4	Bet 5	Bet 6	Bet 7
3 YELLOW	?	?	?	?	?	?	?
85 BROWN	$0.11	$0.05	$0.22	?	?	?	?
7 BLUE	?	?	?	?	?	?	?
5 PURPLE	?	?	?	?	?	?	?

Figure 3. Illustration of the resource-rational SAT-TTB heuristic for multi-alternative risky choice in the Mouselab paradigm where participants choose between bets (red boxes) based on their initially concealed payoffs (gray boxes) for different events (rows) that occur with known probabilities (leftmost column). These payoffs can be uncovered by clicking on corresponding cells of the payoff matrix. The SAT-TTB strategy collects information about the alternatives' payoffs for the most probable outcome (here a brown ball being drawn from the urn) until it encounters a payoff that is high enough (here $0.22). As soon as it finds a single payoff that exceeds its aspiration level, it stops collecting information and chooses the corresponding alternative. The automatic strategy discovery method by Lieder et al. (2017) derived this strategy as the resource-rational heuristic for low-stakes decisions where one outcome is much more probable than all others.

When habits and goal-directed decision-making compete for behavioral control the brain appears to arbitrate between them in a manner consistent with a rational cost–benefit analysis (Daw et al. 2005; Keramati et al. 2011). More recent work has applied the idea of bounded optimality to derive how the habitual and goal-directed decision systems might collaborate (Huys et al. 2015; Keramati et al. 2016). Keramati et al. (2016) found that people adaptively integrate planning and habits according to how much time is available. Similarly, Huys et al. (2015) postulated that people decompose sequential decision problems into subproblems to optimally trade off planning cost savings attained by reusing previous action sequences against the resulting decrease in decision quality.

Overall, the examples reviewed in this section highlight that the principle of resource rationality (Equation 3) provides a unifying framework for a wide range of successful models of seemingly unrelated phenomena and cognitive biases. Resource rationality might thus be able to fill the theoretical vacuum that was left behind by the undoing of expected utility theory. While this section focused on decision-making, the following sections illustrate that the resource-rational framework applies across all domains of cognition and perception.

5. Revisiting classic questions of cognitive psychology

The standard methodology for developing computational models of cognition is to start with a set of component cognitive processes – similarity, attention, and activation – and consider how to assemble them into a structure reproducing human behavior. Resource rationality represents a different approach to cognitive modeling: while the components may be the same, they are put together by finding the optimal solution to a computational problem. This brings advances in AI and ideas from computational-level theories of cognition to bear on cognitive psychology's classic questions about mental representations, cognitive strategies, capacity limits, and the mind's cognitive architecture.

Resource rationality complements the traditional bottom-up approach driven by empirical phenomena with a top-down approach that starts from the computational level of analysis. It leverages computational-level theories to address the problem that cognitive strategies and representations are rarely identifiable from the available behavioral data alone (Anderson 1978) by considering only those mechanisms and representations that realize their function in a resource-rational manner. In addition to helping us uncover cognitive mechanisms, resource-rational analysis also explains why they exist and why they work the way they do. Rational analysis forges a valuable connection between computer science and psychology. Resource-rational analysis strengthens this connection while establishing an additional bridge from psychological constructs to the neural mechanisms implementing them. This connection allows psychological theories to be constrained by our rapidly expanding understanding of the brain.

Below we discuss how resource-rational analysis can shed light on cognitive mechanisms, mental representations, and cognitive architectures, how it links cognitive psychology to other disciplines, and how it contributes to the debate about human rationality.

5.1. Reverse-engineering cognitive mechanisms and mental representations

Resource-rational analysis is a methodology for reverse-engineering the mechanisms and representations of human cognition. This section illustrates the potential of this approach with examples from modeling memory, attention, reasoning, and cognitive control.

5.1.1 Memory

Anderson and Milson's (1989) highly influential rational analysis of memory can be interpreted as the first application of the principle of bounded optimality in cognitive psychology. Their model combines an optimal memory storage mechanism with a resource-rational stopping rule that trades off the cost of continued memory search against its expected benefits (see Equation 3). The storage mechanism presorts memories optimally by exploiting how the probability that a previously encountered piece of information will be needed again depends on the frequency, recency, and pattern of its previous occurrences (Anderson & Schooler 1991), The resulting model correctly predicted the effects of frequency, recency, and spacing of practice on the accuracy and speed of memory recall. While Anderson's rational analysis of memory made only minimal assumptions about its computational constraints, this could be seen as the first iteration of a resource-rational analysis that will be continued by future work.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

More recent research has applied resource-rational analysis to working memory, where computational constraints play a significantly larger role than in long-term memory. For instance, Howes et al. (2016) found that bounded optimality can predict how many items a person chooses to commit to memory from the cost of misremembering, their working memory capacity, and how long it takes to look up forgotten information. Furthermore, resource rationality predicts that working memory should encode information in representations that optimally trade off efficiency with the cost of error (C. R. Sims 2016; C. R. Sims et al. 2012). This optimal encoding, in turn, depends on the statistics of the input distribution and the nature of the task. This allows the model to correctly predict how the precision of working memory representations depends on the number of items to be remembered and the variability of their features. Over time working memory also have to dynamically reallocate its limited capacity across multiple memory traces depending on their current strength and importance (Suchow 2014). Suchow and Griffiths (2016) found that the optimal solution to this problem captured three directed remembering phenomena from the literature on visual working memory.

5.1.2. Attention

The allocation of attention allows us to cope with a world filled with vastly more information than we can possibly process. Applying resource-rational analysis to problems where the amount of incoming data exceeds the cognitive system's processing capacity might thus be a promising approach to discovering candidate mechanisms of attention. Above we have reviewed a number of bounded optimal models of the effect of limited attention on decision-making (Caplin & Dean 2015; Caplin et al. 2017; Gabaix 2014; 2016; 2017; Lieder 2018; C. A. Sims 2003; 2006), so this section briefly reviews resource-rational models of visual attention.

The function of visual attention can be formalized as a decision-problem in the framework of partially observable Markov decision processes (POMDPs; Gottlieb et al. 2013) or meta-level Markov decision processes (Lieder et al. 2017). Such decision-theoretic models make it possible to derive optimal attentional mechanisms. For instance, Lewis et al. (2014) and Butko and Movellan (2008) developed bounded optimal models of how long people look at a given stimulus and where they will look next, respectively, and the resource-rational model by Lieder et al. (2018e) captures how visual attention is shaped by learning.

Finally, resource-rational analysis can also elucidate how people distribute their limited attentional resources among multiple internal representations and how much attention they invest in total (Van den Berg & Ma 2018). Among other phenomena, the rational deployment of limited attentional resources can explain how people's visual search performance deteriorates with the number of items they must inspect in parallel. To explain such phenomena the model by van den Berg and Ma (2018) assumes that the total amount of attentional resources people invest is chosen according to a rational cost–benefit analysis that evaluates the expected benefits of allocating more attentional resources against their neural costs (see Equation 3).

5.1.3. Reasoning

Studies reporting that people appear to make systematic errors in simple reasoning tasks (e.g., Tversky & Kahneman 1974; Wason 1968) have painted a bleak picture of the human mind that is in stark contrast to impressive human performance in complex problems of vision, intuitive physics, and social cognition. Taking into account the cognitive constraints that require people to approximate Bayesian reasoning might resolve this apparent contradiction (Sanborn & Chater 2016), and resource-rational analyses of how people overcome the computational challenges of reasoning might uncover their heuristics (e.g., Lieder et al. 2018a; 2018b).

One fundamental reasoning challenge is the frame problem (Fodor 1987; Glymour 1987): Given that everything could be related to everything, how do people decide which subset of their knowledge to take into account for reasoning about a question of interest? The resource-rational framework can be applied to derive which variables should be considered and which should be ignored depending on the problem to be solved, the resources available, and their costs. In an analysis of this problem, Icard and Goodman (2015) showed that it is often resource-rational to ignore all but the one to three most relevant variables. Their analysis explained why people neglect alternative causes more frequently in predictive reasoning ("*What will happen if …*") than in diagnostic reasoning ("*Why did this happen?*"). Nobandegani and Psaromiligkos (2017) extended Icard and Goodman's analysis of the frame problem toward a process model of how people simultaneously retrieve relevant causal factors from memory and reason over the mental model constructed thus far. Future work should extend this approach to studying alternative ways in which people simplify the mental model they use for reasoning and how they select this simplification depending on the inference they are trying to draw and their reasoning strategy.

Recently, the frame problem has also been studied in the context of decision-making (Gabaix 2014; 2016). Gabaix's characterization of a resource-rational solution to this problem predicts many systematic errors in human reasoning, including base-rate neglect, insensitivity to sample size, overconfidence, projection bias (the tendency to underappreciate how different the future will be from the present), and misconceptions of regression to the mean (Gabaix 2017).

Resource-rational analysis has also already shed light on two additional questions about human reasoning: "How do we decide how much to think?" and "From where do hypotheses come?" Previous research on reasoning suggested that people generally think too little, a view that emerged from findings such as the anchoring bias (Tversky & Kahneman 1974), according to which people's numerical estimates are biased toward their initial guesses (Epley & Gilovich 2004). Contrary to the traditional interpretation that people think too little, a resource-rational analysis of numerical estimation suggested that many anchoring biases are consistent with people choosing the number of adjustments they make to their initial guess in accordance with the optimal speed-accuracy trade-off defined in Equation 3 (Lieder et al. 2018c; 2018d). Drawing inspiration from computer science and statistics, this resource-rational analysis yielded a general reasoning mechanism that iteratively proposes adjustments to an initial idea; the proposed adjustments are probabilistically accepted or rejected in such a way that the resulting train of thought eventually converges to the Bayes-optimal inference.

The idea that people generate hypotheses in this way can explain a wide range of biases in probabilistic reasoning (Dasgupta et al. 2017) and has since been successfully applied to model how people reason about causal structures (Bramley et al. 2017), medical diagnoses, and natural scenes (Dasgupta et al. 2017; 2018). A subsequent resource-rational analysis revealed that once people have generated a hypothesis in this way they memorize it and later retrieve it to more efficiently reason about related questions in the future (Dasgupta et al. 2018).

5.1.4. Goals, executive functions, and mental effort

Goals and goal-directed behavior and cognition are essential features of the human mind (Carver & Scheier 2001). Yet, from the perspective of expected utility theory (Equation 1), there is no reason why people should have goals in the first place. An unboundedly optimal agent would simply maximize its expected utility by scoring all outcomes its actions might have according to its graded utility function. In contrast, people often think only about which subgoal to pursue next and how to achieve it (Newell & Simon 1972). This is suboptimal from the perspective of expected utility theory, even though it seems intuitively rational for people to be goal-directed, and empirical studies have found that setting goals and planning how to achieve them is highly beneficial (Locke & Latham 2002). The resource rationality framework can reconcile this tension by pointing out that goal-directed planning affords many computational simplifications that make good decision-making tractable. For instance, planning backward from the goal – as in means-ends analysis (Newell & Simon 1972) – allows decision-makers to save substantial amounts of computation by ignoring the vast majority of all possible states and action sequences. Future work will apply resource rationality to provide a normative justification for the existence of goals and develop an optimal theory of goal-setting.

Our executive functions adapt and organize how we think and decide to the goals we are currently pursuing; without them, our thoughts would be incoherent and our behavior disorganized, and we would be unable to achieve even our most basic objectives. Executive functions are effectively the mechanisms through which goals enable us to reason and act effectively in the face of complexity that exceeds our cognitive capacities. To achieve resource rationality, cognitive control should be allocated in accordance with a rational cost–benefit analysis that weights improved performance against the time, effort, and cognitive resource costs needed to achieve it (Shenhav et al. 2013; Shenhav et al. 2017; see Equation 3). Encouragingly, resource rationality has already shed light on how control is allocated between alternative cognitive mechanisms (Lieder & Griffiths 2017; Shenhav et al. 2013) and decision systems (Boureau et al. 2015; Daw et al. 2005; Keramati et al. 2011). Furthermore, it can explain how much mental effort people exert (Dickhaut et al. 2009; Shenhav et al. 2017), whether and how intensely competing automatic processes will be inhibited (Lieder et al. 2018e), how people can flexibly switch between alternative strategies (Lieder & Griffiths 2017; Payne et al. 1993), and people's occasional lapses of self-control (Boureau et al. 2015).

5.1.5. Mental representations

How does the mind encode information and how does it structure our knowledge about the world around us? While the principle of bounded optimality was originally formulated for programs and has been predominantly applied to model cognitive strategies, it can also be applied to model mental representations. There are already several successful applications of bounded optimality to modeling perceptual representations, representations in visual working memory, representations of decision variables, task representations, and the way we use language to represent the world. In our discussion of the frame problem and decision-making with limited attentional resources, we already saw that bounded optimality can shed light on which features and variables should and shouldn't be included in mental representations (Gabaix 2014; 2016; Icard & Goodman 2015). Here, we focus on how the attended features of the environment should be represented.

From a Bayesian perspective people should leverage their prior knowledge about the statistics of the world to resolve perceptual uncertainty. For instance, people should resolve their uncertainty about the exact orientation of a line in favor of the more common orientation and thus be more likely to perceive an almost vertical line to be closer to vertical than farther from vertical. But curiously it is just the opposite. Wei and Stocker (2015; 2017) showed that the optimal allocation of limited representational resources across different stimulus features can explain this puzzling perceptual bias that distorts our perception of the world away from what we should expect to see. This illustrates that apparently irrational perceptual illusions can arise from bounded-optimal information processing. Polania et al. (2019) found that the same principles also predict how the biases and variability in how people judge the value of consumer products and choose among them depends on the products' value.

Resource-rational analysis can also elucidate the format of mental representations. For instance, Bhui and Gershman (2017) derived that the brain should represent utilities and probabilities by their smoothed rank (e.g., representing "$500" as "more expensive than about 75% of the products in this category"). This representation explains why people's preferences often violate the prescriptions of expected utility theory (Stewart 2009; Stewart et al. 2006).

While the model by Bhui and Gershman (2017) specifies the representation of numeric quantities, information theoretic models developed by Chris R. Sims and colleagues implicitly define resource-rational perceptual representations that are optimized for making good decisions in the face of capacity constraints and noise. Specifically, they use rate-distortion theory to show that perception and working memory should encode information in representations that optimally trade off their efficiency versus the cost of error to explain the limitations of human performance in absolute identification (where the task is to report to which of n taught categories each stimulus belongs) and visual working memory (C. R. Sims 2016; C. R. Sims et al. 2012). This approach emphasizes that representations are shaped by the behavioral consequences of perceptual errors; for instance, consistent with error management theory (Haselton & Nettle 2006), our representations should reflect that it is much costlier to misperceive a poisonous mushroom as edible than to confuse two edible mushrooms.

Similar information-theoretic principles have also been applied in the domain of language (Hawkins 2004; Kemp & Regier 2012; Regier et al. 2007; Zaslavsky et al. 2018; Zipf 1949). According to Zipf's *principle of least effort*, speakers aim to communicate their message with as little effort as possible while still being understood by the listener (Zipf 1949). This principle has been successfully applied to explain why the frequency of a word is inversely proportional to its rank (Zipf 1949) and why some words are shorter than others (Mahowald et al. 2013; Piantadosi et al. 2011; Zipf 1949). Similar effort-accuracy tradeoffs can also explain how people represent colors (Regier et al. 2007; Zaslavsky et al. 2018) and kinship relations (Kemp & Regier 2012) and could potentially be invoked to understand chunking (Gobet et al. 2001) as a bounded-optimal mechanism for representing information in memory to reduce the cost of memory maintenance while increasing recall performance.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Future resource-rational analyses might elucidate many additional representations. For instance, the principle of resource rationality could be applied to derive hierarchical action representations (Bacon et al. 2017; Botvinick 2008; Solway et al. 2014) that achieve the best possible trade-off between planning efficiency and reduced behavioral flexibility.

5.2. Cognitive architectures and capacity limits

Resource-rational models can also be used to revisit some of cognitive psychology's foundational debates about the nature of the mind's cognitive architecture, its potential subsystems (which might, among others, include declarative memory, procedural memory, the visual system, and the central executive), and their capacity constraints (e.g., Lewis et al. 2014; C. R. Sims 2016; C. R. Sims et al. 2012; van den Berg & Ma 2018). Resource-rational analysis has already led to a fundamental rethinking of the limits of working memory (C. R. Sims 2016; C. R. Sims et al. 2012; Van den Berg & Ma 2018), attention (Van den Berg & Ma 2018), and cognitive control (Howes et al. 2009; Musslick et al. 2016; Segev et al. 2018), and it is beginning to elucidate why the mind appears to be structured into a small number of subsystems (Milli et al. 2017; 2019).

C. R. Sims et al. (2012) used resource-rational modeling to translate alternative assumptions about the capacity limits of visual working memory into quantitative predictions. Testing these predictions against empirical data suggested that visual working capacity is not limited to a fixed number of items but can be flexibly divided to store either a small number of items with high fidelity or a larger number of items with lower fidelity. This approach also suggested that people's working memory capacity may be higher than currently assumed because people's performance in working memory tasks may be limited by unnatural stimulus statistics (Orhan et al. 2014). Taking this approach even further, van den Berg and Ma (2018) have recently challenged the engrained assumption that working memory always distributes a *fixed amount* of representational resources among the encoded items by showing that the effect of working memory load on performance is better explained by a mechanism that adjusts the total amount of working memory resources according to a rational cost–benefit analysis.

Another classic debate in cognitive psychology concerned the question of serial processing (e.g., Sternberg 1966) versus parallel information processing (Atkinson et al. 1969) in perception, short-term memory, attention (Eckstein 1998; Treisman & Gelade 1980; Wolfe 1994) and multitasking (Fischer & Plessow 2015). Recent applications of bounded optimality revealed that resource-constrained parallel processing can produce effects that look like serial processing (Howes et al. 2009, Musslick et al. 2016; 2017, Segev et al. 2018).

While some have argued that the capacity limits in multitasking arise from a single, capacity-limited, serial-processing mechanism (Anderson et al. 2004; Pashler & Sutherland 1998), recent resource-rational analyses (Feng et al. 2014; Musslick et al. 2016) supports the alternative view that capacity limits for multitasking reflect parallel processes competing for limited local resources (Allport et al. 1972; Meyer & Kieras 1997a; 1997b). The bottleneck that the neural pathways of different functions compete for shared representations may itself be a consequence of the rational use of limited resources because shared representations support faster learning through generalization (Musslick et al. 2017; Segev et al. 2018).

More generally, this illustrates that applying the principle of bounded optimality to the design of cognitive systems can explain why certain cognitive limitations exist at all. It is conceivable that other cognitive limits also arise from a rational trade-off between the capacity to learn highly specialized, maximally performant cognitive mechanisms and the amount of time and experience that this would require.

Finally, the resource-rational approach can also be used to derive optimal cognitive architectures (Milli et al. 2017; 2019), thereby generating principled hypotheses about how, which, and how many cognitive systems the mind should be equipped with. Empirically testing the predictions of such models, revising their assumptions accordingly, re-deriving the optimal cognitive architecture, and then repeating this process until the predictions are sufficiently accurate extends resource-rational analysis from reverse-engineering cognitive mechanisms to reverse-engineering cognitive architectures. Milli et al. (2017; 2019) found that this methodology can provide a resource-rational justification for the apparent prevalence of the coexistence of fast but error-prone sub-systems with slow but accurate sub-systems in human reasoning (Evans 2008; Stanovich 2011), judgment (Kahneman & Frederick 2002; 2005), and decision-making (Dolan & Dayan 2013).

5.3. Connecting psychology to AI and neuroscience

Neuroscience, psychology, economics, and AI investigate intelligence and decision-making at different levels of abstraction. Neuroscience takes the brain's anatomical, physiological, and biophysical constraints very seriously. Psychology works with abstract models of the mind that ignore many of the brain's computational constraints. And economics and AI research simplify and idealize these models of the mind even further. Resource-rational analysis connects these different levels of abstraction by taking an abstract model of the mind of the kind that might be developed in economics and AI research and augments it with increasingly more realistic psychological and/or neurobiological constraints. In doing so, resource-rational analysis establishes new bridges between these various disciplines (see Fig. 4).

5.3.1 Connecting levels of analysis: Case studies from perception and efficient coding

The iterative refinements that resource-rational analysis makes to its assumptions about the mind's cognitive architecture (see Box 2) generally proceed from the most abstract and most unconstrained model of the underlying neurocognitive architecture (see Fig. 4). Resource-rational analysis builds bridges from the computational level of analysis to the algorithmic level and then the implementational level. In this way, models of cognitive strategies and representations can be informed by both theories of AI and biophysical constraints on computation and representation.

The application of resource rationality to Marr's implementation level and its connection to the algorithmic level has been most thoroughly explored in the domain of perception. Bounded-optimal models of perception generally assume that the brain receives too much sensory input to represent all of it accurately and that the accuracy of a neural representation is limited by how much neural resources have been allocated to it. Bounded optimality has been applied to both the allocation of neural resources (Ganguli & Simoncelli 2014; Wei & Stocker 2015; 2017) and the use of the resulting noisy representations (Stocker et al. 2006).

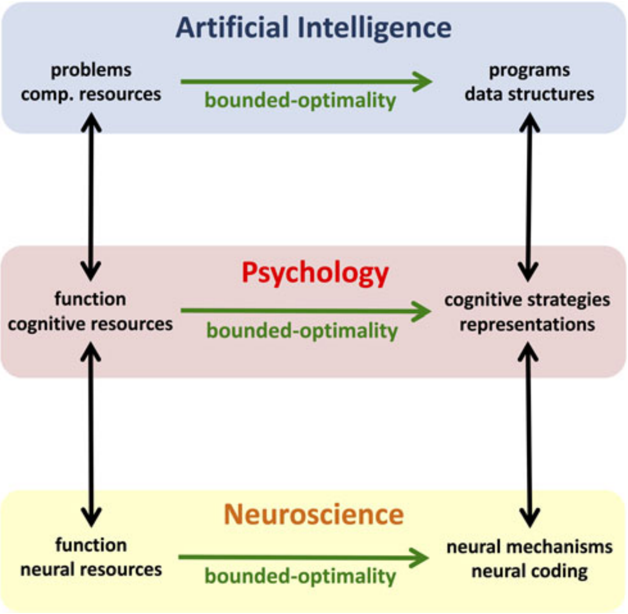


Figure 4. Resource-rational analysis connects levels of analysis.

The principles of resource rationality can also be applied to understand how neural mechanisms of perception are shaped by metabolic and biophysical constraints. For instance, action potentials are so metabolically expensive that only about 1% of all neurons in the brain can sustain substantial activity in parallel (Lennie 2003). This limit imposes serious constraints on how the brain can represent and process information, and many aspects of morphology, physiology, and the wiring of neural circuits can be understood as adaptation to the evolutionary pressure to achieve a near-optimal trade-off between computational efficacy and metabolic cost (Levy & Baxter 2002; Niven & Laughlin 2008; Sterling & Laughlin 2015). This principle can be applied to derive neural codes that encode as much information as possible with as little neural activity as necessary (Levy & Baxter 1996; Olshausen & Field 1996; 1997; 2004; Wang et al. 2016). Another success story where bounded optimality assisted in connecting the algorithmic level of analysis to the implementation level are the neural inhibition models of optimal perceptual decision-making (Bogacz et al. 2006; Van Ravenzwaaij et al. 2012). Finally, the effects of metabolic constraints are not restricted to details of the neural implementation but propagate all the way up to high-level cognition by necessitating cognitive mechanisms like selective attention (Lennie 2003).

5.3.2 Transfer of ideas between computer science and cognitive science

Another key advantage of bounded optimality is that it provides a common language for computer science, psychology, and neuroscience researchers to exchange ideas across disciplinary boundaries. There are already many examples of cognitive models inspired by ideas from computer science in general and optimal algorithms in particular (Anderson 1990; Gershman et al. 2015; Griffiths et al. 2012; 2015; Sanborn et al. 2010). Some key AI advances have been inspired by neuroscience and psychology (Hassabis et al. 2017), reinforcement learning and deep learning being prime examples.

Under the assumption that the brain is approximately bounded-optimal, the endeavor to uncover people's cognitive strategies and representations becomes a pursuit of optimal algorithms and data structures for problems such as inference, learning, control, and decision-making. Discovering such algorithms is the long-standing goal of AI. Computational efficiency has always been a key objective in computer science, and research in AI, robotics, and machine learning is increasingly tackling the hard problems of perception, learning, motor control, and reasoning that people solve daily. Thus, AI research on bounded optimality can be expected to provide continued inspiration for uncovering how the mind works. One way to encourage more AI research on bounded optimality could be to introduce new benchmark tasks that explicitly limit the computational resources used to solve the problem to a biologically plausible level.

Conversely, as the paradigm of bounded optimality orients psychology and neuroscience toward the computational mechanisms through which the brain achieves its tremendous computational efficiency, the resulting insights will likely to continue to inspire advances in AI (Lake et al. 2017; Nobandegani 2017).

5.4. Rationality revisited

Research is now revisiting the debate about human rationality with resource rationality as a more realistic normative standard. The results are beginning to suggest that heuristic mechanisms that are commonly interpreted as evidence against human rationality might not be irrational after all. Instead, they might reflect the optimal use of finite time and limited computational resources. For instance, the tendency to over-estimate the frequency of extremely good and extremely bad events and to overweight them in decision-making might reflect a bounded optimal decision mechanism that prioritizes the most important eventualities (Lieder et al. 2018b). In addition, the principle of resource rationality can also explain contextual preference reversals (Howes et al. 2016), risk aversion (Khaw et al. 2017), wishful thinking (Neuman Rafferty & Griffiths 2014), sub- and super-additive biases in probability judgments (Dasgupta et al. 2017; 2018), perceptual biases (Stocker et al. 2006; Wei & Stocker 2015; 2017), hyperbolic discounting, base rate neglect, the law of small numbers, and many more, including the probability distortions described by prospect theory (Gabaix 2017).

These findings collectively suggest that the interpretation of cognitive biases as a sign of human irrationality must be reconsidered – it is too early to conclude that people are fundamentally irrational (Ariely 2009; Marcus 2008; Sutherland 2013). Instead, a valid answer to the question of human rationality will require thorough evaluations of human cognition against the predictions of resource rationality (Equation 4). This perspective also suggests that we should redefine the term "cognitive bias" as a violation of resource rationality rather than as a violation of logic, probability theory, or expected utility theory.

As reviewed above, resource-rational analysis can rationalize some cognitive biases as a consequence of certain capacity limits. But for people's heuristics to be considered truly resource-rational, it is not enough for them to be optimal with respect to some *hypothetical* cognitive constraints; to be resource-rational people's heuristics have to be optimal with respect to their *actual* cognitive constraints. This makes independently measuring people's cognitive constraints an important direction for future work. If people's heuristics turned out to be optimal relative to their cognitive limitations, then one might subsequently ask "Is

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

it rational for people's cognitive capacities to be so limited or should evolution have equipped us with better brains?". This question could be addressed by performing cost–benefit analyses similar to those defined in Equation 4 to determine to which extent evolution has succeeded to design resource-rational neural hardware (Sterling & Laughlin 2015). If we were able to derive what people's cognitive capacities should be, this would provide a very principled starting point for resource-rational analysis.

5.4.1 Implications for improving the human mind

In addition to its contributions to understanding the human mind, resource rationality also provides guidance for how to improve it. These prescriptions are fundamentally different from the standard approach of debiasing (Larrick 2004) that aims to reduce or eliminate people's deviations from the rules of logic, probability theory, and expected utility theory – usually by educating people about these rational principles or their implications. Instead, the resource-rational perspective suggests that people should be taught simple heuristics that make optimal use of their limited cognitive resources. Recent technical advances (Callaway et al. 2018a; Lieder Krueger & Griffiths 2017) make it possible to discover and teach resource-rational heuristics automatically (Lieder et al. 2018a; Sedlmeier & Gigerenzer 2001). Alternatively, resource constraints could be addressed through cognitive training or cognitive prostheses like navigation systems or decision-support systems (e.g., Lieder et al. 2019b).

Resource-rational analysis can also help us decide which interventions are most appropriate for improving performance. For instance, a resource-rational analysis of a person's scores on a series of tests could reveal that their performance is primarily limited by verbal working memory, in which case working memory training might be effective. In other situations, people's inferences or decisions might indeed be rational under reasonable assumptions about the structure of the environment that are violated by the current situation. In these cases, the prescription might be to align the presentation of such problems with the implicit assumptions of the strategies that people use to solve them.

6. Challenges of resource-rational analysis

Having illustrated the potential of resource-rational analysis, we now turn to its limitations and challenges: scenarios where the prerequisites of resource-rational analysis may not hold, people's apparent irrationality, knowing what the cognitive constraints are, testing resource-rational models empirically, and applying resource-rational analysis to the real-world.

Resource-rational analysis is predicated on the assumption that cognitive mechanisms are well-adapted to their function and the cognitive constraints under which they operate. Adaptation can be achieved through evolution or learning. For evolutionary arguments to hold, the evolutionary environment must have exerted sufficiently strong adaptive pressures over sufficiently long periods of time and the assumptions about the evolutionary environment must be accurate. And adaptation through learning requires a sufficient amount of relevant experience. Cases where these assumptions are violated or difficult to specify are challenging for resource-rational analysis. This includes people's performance during the process of adaptation to a new environment and infrequent situations where people's performance has no critical ramifications. Resource-rational analysis is especially difficult to apply when the environment or cognitive constraints are unknown. Furthermore, adaptive pressures constrain cognitive mechanisms only to the extent that performance is sensitive to changes in the mechanism. Thus, if there is a wide range of different mechanisms that perform almost equally well, then the outcome of adaptation need not be resource-rational.

Everyday observations of seemingly irrational beliefs and behaviors and empirical demonstrations of cognitive biases constantly challenge the view that people are resource-rational. As reviewed above, people's decision-mechanisms appear to be surprisingly resource-rational. But even when people believe they understand something deeply their intuitive theories are often shallow and fragmented (Rozenblit & Keil 2002). This apparent contradiction dissolves in scenarios where irrational beliefs do not manifest in perilous decisions with costly consequences. The adaptive pressures that mold decision mechanisms into a resource-rational shape do not apply to how people learn and reason about X (e.g., astronomy or philosophy) if their beliefs about X have little effect on the decisions determining their evolutionary fitness and the rewards they learn from (cf. Equation 2). In such cases, having questionable beliefs about X is not inconsistent with being (approximately) resource-rational. To the contrary, to be resource-rational the mechanisms of cognitive capacities that are far removed from important decisions should be extremely efficient even at the expense of their accuracy.

Identifying and quantifying the resource constraints on cognitive mechanisms and representations can be very challenging. Ideally, such assumptions should be grounded in independent measurements of cognitive capacities, such as processing speed or working memory capacity, or biological constraints, such as nerve conduction velocity, metabolic constraints on the amount of simultaneous neural activity, or the maximum rate at which a neuron can fire. Only when such constraints have been established empirically, can we interpret the resulting resource-rational heuristic as a normative standard for human reasoning or decision-making. But in practice cognitive constraints often have to be estimated through parameter fitting and model comparison.

Encouraging modelers to revise their assumptions about cognitive constraints in the face of data (i.e., step 5 in Box 2) is a double-edged sword. It can be useful to generate hypotheses about the mind's capacity limitations and to find good explanations of otherwise puzzling phenomena. But postulating cognitive constraints carelessly without good theoretical and empirical reasons could also produce bad models that overfit observations of idiosyncratic or genuinely irrational behaviors with wrong assumptions. To guard against this, one should ideally base all assumptions about the constraints on independent empirical measurements. Assumptions about biological constraints can be derived from physiological measurements and assumptions about cognitive constraints can, at least in principle, be derived from psychometric tests that isolate the capacity of interest and ensure that people are motivated to perform as well as possible. When the unavailability of such measurements makes it necessary to resort to assumptions and parameter estimation, then the resulting resource-rational model should not be evaluated by its fit to the modelled data set but by its ability to predict other phenomena that it was not designed to capture, and the model's assumptions about resource constraints should be empirically tested in subsequent research. The fact that capacity constraints are real, measurable properties of the brain makes resource-rational models falsifiable. But we acknowledge that, to date, measuring cognitive constraints remains challenging and often requires additional assumptions. The resulting uncertainty

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

about people's cognitive constraints can make it challenging to falsify resource-rational models in practice. This makes measuring cognitive capacities, such as the speed with which various elementary cognitive operations can be performed, an important direction for future work.

Applying rational principles to modeling higher-level cognition is controversial because many researchers believe that the heuristics that resource-rational analysis is meant to uncover are arbitrary and irrational (Ariely 2009; Gilovich et al. 2002; Marcus 2008) and call for different organizing principles (e.g., Kahneman 2003) such as evolutionary history (e.g., Buss 1995; Marcus 2008; Todd & Gigerenzer 2012). We have argued that evolutionary adaptation might have molded the mind into a roughly resource-rational shape. But since evolution does not necessarily produce optimally adapted phenotypes some argue that heuristics are kluges that can only be understood as accidents of evolutionary history (Marcus 2008). Our framework partially accounts for evolutionary history by considering that cognitive mechanisms may be adapted to a mixture of different environments (Equation 4) – potentially including a series of past evolutionary environments. Other researchers may argue that mathematical theories of brain function, such as the free-energy principle (Friston 2010), provide a more appropriate theoretical framework for understanding the mechanisms of perception, learning, and decision-making than our notion of resource rationality. Finally, it is conceivable that theoretical constraints will become less important to cognitive modeling as we get more data and increasingly more refined methodologies for measuring the neurocognitive mechanisms of reasoning and decision-making. But in our view, resource-rational analysis is a very promising methodology and time will tell under which conditions its methodological assumptions are useful.

So far, resource-rational modeling and automatic methods for discovering and teaching rational heuristics have only been applied to laboratory paradigms whose structure is simple and fully known. It will be challenging to scale these approaches to decision-making in the real world where the sets of options and possible outcomes are much larger and often unknown. Equation 4 provides a theoretical framework for incorporating such uncertainties into the design of heuristics that are robust to errors in our models of the environment. This robustness is achieved by optimizing the heuristic's average performance across all environments that are consistent with our limited knowledge (weighted by their likelihood), and recently developed methods for discovering optimal heuristics (Callaway et al. 2018a; in preparation) can already handle this formulation of uncertainty about the environment. Future work should also continue to measure the structure of natural decision environments because the heuristics our methods discover will only be as good as our models of the problems they are meant to solve. Good models of people's cognitive constraints and robustness to their imperfections are equally critical – especially for improving human performance. For instance, a memory strategy optimized for a working memory span of 7 items, might be disastrous for a person who can hold only 4 items in memory. Future work will therefore incorporate uncertainty about people's cognitive capacities into the definition of rational heuristics in the same way as Equation 4 incorporates uncertainty about the environment. The ultimate criterion for the rationality of automatically discovered heuristic will be how well people perform when they use them in the real world.

7. Conclusion

Resource-rational analysis is an emerging paradigm for modeling human cognition that leverages bounded optimality to simultaneously explain both people's seemingly irrational cognitive biases and their remarkable capacity to solve almost effortlessly complex problems that continue to elude AI. This approach integrates the strengths of rational theories with the psychological realism of descriptive models of cognitive mechanisms and representations. The studies reviewed above illustrate that resource rationality provides a unifying principle for answering fundamental questions about perception, decision-making, memory, attention, reasoning, and cognitive control. This unifying framework can be used to build bridges between psychology, neuroscience, AI, and economics (see Fig. 4). Furthermore, resource rationality also allows us to answer teleological questions about the nature of the mind, such as why we represent and think about the world the way we do, what the purpose of goals is, and why the mind is divided into a small number of modular subsystems. Finally, by enabling the development of quantitative benchmarks of bounded rationality, resource-rational analysis sheds new light on the debate about human rationality and opens new avenues to improving the mind.

Although the idea that the mind strives to maximize utility under cognitive constraints has been around for a long time, the systematic, quantitative methodology of resource-rational analysis is a recent development and much more work remains to be done to strengthen its foundation and establish it as a new paradigm for cognitive modeling. Resource-rational models could be made substantially stronger by grounding them in increasingly realistic assumptions about the brain's computational architecture and capacity limits. To achieve this, future work should integrate resource-rational analysis with previous work on cognitive architectures and establish a solid empirical foundation for its assumptions about capacity limits and computational costs. Measuring the bounds on human cognition will permit rigorously testing the methodological assumption that people make rational use of their limited cognitive resources. This line of research will help establish to what extent resource-rational models are psychologically plausible. At best, resource rationality could become a principled methodology for discovering people's cognitive mechanisms and representations from the biophysical limits on neural information processing. At worst, resource rationality could turn out to be a convenient template for slightly less unrealistic as-if explanations than standard models based on Bayesian inference and expected utility theory.

Recent work suggests that the assumption of resource rationality becomes increasingly accurate as people continue to learn about and adapt to a new environment (e.g., Lieder & Griffiths 2017). Learning how to make rational use of limited resources may be an essential component of cognitive development and a necessity for adapting to evolving environments. We therefore believe that a complete theory of resource rationality needs to include a bounded-optimal mechanism for learning to become resource-rational. We are currently investigating this learning mechanism by studying how people learn how to think and decide.

We hope that resource-rational analysis will mature into a widely used paradigm for elucidating the mechanisms of human cognition with mathematical precision. In addition to its contributions to reverse-engineering cognitive mechanisms, bounded optimality might also advance psychological research much the way classic notions of rationality gave rise to the blooming field of judgment and decision-making: by providing a normative standard against which human performance can be compared to characterize in which ways people's heuristics deviate from resource-rational

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

strategies. However, since bounded optimality provides a much more realistic normative standard than did expected utility theory, logic, and probability theory, we might find that our minds are much more rational than we thought. We still have a long way to go but, in our view, resource rationality is a promising framework for modeling the human mind with mathematical precision.

Acknowledgments. We would like to thank Florian Mohnert, Sayan Gul, Fred Callaway, Charles Kozierok, Ardavan Nobandegani, Daniel Reichman, and Rachit Dubey for helpful comments on an earlier version of this article. This research was funded under grant number ONR MURI N00014-13-1-0341 from the Office of Naval Research, contract FA9550-18-1-0077 from the Air Force Office of Scientific Research, and a grant from the Templeton World Charity Foundation.

Open Peer Commentary

What are the appropriate axioms of rationality for reasoning under uncertainty with resource-constrained systems?

Harald Atmanspacher[a], Irina Basieva[b], Jerome R. Busemeyer[c], Andrei Y. Khrennikov[d], Emmanuel M. Pothos[b], Richard M. Shiffrin[c] and Zheng Wang[e]

[a]Collegium Helveticum, Zürich, 8006 Switzerland; [b]Department of Psychology, City University London, London EC1V 0HB, United Kingdom; [c]Psychological Brain Sciences, Indiana University, IN 47405; [d]Department of Mathematics at Linnaeus University, Linnaeus University, 351 95 Växjö, Sweden; and [e]Department of Communication, The Ohio State University, Columbus, OH 43210

atmanspacher@collegium.ethz.ch
https://collegium.ethz.ch/en/about-us/staff/pd-dr-harald-atmanspacher/
irina.basieva@gmail.com
https://uk.linkedin.com/in/irina-basieva-3182b1108
jbusemey@indiana.edu
http://mypage.iu.edu/~jbusemey/home.html
andrei.khrennikov@lnu.se
https://lnu.se/en/staff/andrei.khrennikov/
Emmanuel.Pothos.1@city.ac.uk
https://www.city.ac.uk/people/academics/emmanuel-pothos
shiffrin@indiana.edu
http://shiffrin.cogs.indiana.edu
wang.1243@osu.edu
https://comm.osu.edu/people/wang.1243

doi:10.1017/S0140525X19001535, e2

Abstract

When constrained by limited resources, how do we choose axioms of rationality? The target article relies on Bayesian reasoning that encounter serious *tractability* problems. We propose another axiomatic foundation: quantum probability theory, which provides for less complex and more comprehensive descriptions. More generally, defining rationality in terms of axiomatic systems misses a key issue: rationality must be defined by humans facing vague information.

The main thesis of the target article is that the mind is based on a rational use of limited resources. We agree that this is a useful organizing principle as long as we interpret "rational reasoning" as deriving from coherent axioms. However, when the mind is constrained by limited resources, the issue of how best to choose axioms of rationality becomes a matter of debate. In particular, the target article relies heavily on Bayesian reasoning tools that encounter serious *tractability* problems. This is because the dimension of the probability space grows exponentially out of control as the number of variables increases. This is a well-known problem recognized by the proponents of Bayesian cognition (e.g. Tenenbaum et al. 2011). Consequently, resource limited *extensions beyond* basic Bayesian reasoning are required that rely on various approximations for simplifying computations, for example, through sampling approximations (Sanborn et al. 2010) and/or employing Bayesian networks to truncate complex conditional dependencies (Lake et al. 2015). Are these approximations really resource rational? And are these the only ways to meet the resource constraints for reasoning under uncertainty?

We propose another resource rational alternative, where, as above, rational status is justified by an axiomatic foundation: quantum probability theory (e.g. Aerts et al. 2013; Basieva et al. 2018; Bruza et al. 2015; Khrennikov et al. 2018; Pothos & Busemeyer 2013; Wang et al. 2014; Yukalov & Sornette 2011). One advantage of quantum probability theory is that it provides more parsimonious (less complex) descriptions than Bayesian approaches based on Kolmogorov probability theory (Atmanspacher & Römer 2012). The dimension of the probability space does not increase exponentially, and in certain circumstances, it does not increase with increasing number of variables. How does this work?

Kolmogorov probability theory (which forms the basis of Bayesian theory) is founded on assignment of probabilities to events represented subsets of a sample space, which assumes a complete Boolean algebra of events. Quantum theory assigns probabilities to measurement outcomes, represented as subspaces of a vector space, which entails only a partial Boolean algebra. A theorem by Gleason (1957) states that any additive measure used to assign probabilities to subspaces of a vector space (with dimension greater than 2) can be described as quantum probabilities. The non-Boolean aspect of quantum theory arises from the use of non-commutative observables, which implies sequence effects for the results of successive measurements. Wang et al. (2014) demonstrated convincingly how powerful quantum modeling proves to be in this regard.

The advantage of using a vector space representation is that different measurements can be described by changing the basis used to define them. There is an infinite number of ways to select a basis within a fixed and finite vector space, which can then provide an infinite number of ways to describe concepts within a limited cognitive resource. An example will help illustrate this important point. Consider a game with two players, and each player has three moves. When planning a move, each player needs to estimate the probability of the move of the opponent and then consider the probability for his/her own move. According to a Bayesian probability model, this requires forming $3 \times 3 = 3^2$ joint probabilities that each of two players takes one of three actions. If there are n players, then a Bayesian model requires 3^n joint probabilities, producing an exponential growth in probabilities. In contrast, according to the quantum approach, the state of the three actions by each player can be represented by a unit length vector in a 3-dimensional space. The probabilities assigned to different players can be obtained by "rotating" the basis used to describe the vector within the same 3-dimensional space. In this way, n players are described by n different bases within the same 3-dimensional space.

There is cognitive cost produced by representing different measurements using different bases which is expressed by a quantum-like uncertainty principle. In our n-person game example, it is not possible to be certain about the moves of all players simultaneously. Increasing certainty about the move of one player implies increasing uncertainty about others. In quantum physics, the uncertainty principle is a consequence of the structure of the physical world; for psychology, we propose that its relation to limited cognitive resources may be a structural feature of the mental world.

Which approach to forming a rational reasoning system under uncertainty is most appropriate? Partly, this is a computational problem (i.e. which approach provides the optimal balance between precision and simplicity), and partly this is an empirical problem (i.e. which approach predicts better apparent inconsistencies/errors in human judgments).

Going beyond competing axiomatic reasoning systems, a broader issue really needs to be addressed. It is natural to attempt to characterize and *quantify* human limited cognitive resources, and then to argue that the decisions are optimal in light of corresponding limitations. However, humans have evolved to make decisions when quantification is impossible. When one quantifies cognition, one specifies uncertainty in terms of specifying distributions of all and any variables in the system. This idea lies at the heart of the resource rational analysis. Yet most information in life is vague and defies quantification. When, for example, we must decide who to marry, which job offer to accept, what house to buy, or any of the normal decisions human face, we cannot specify the relevant distributions in any way we trust. This lack of precision does not mean we have no useful information – there is almost always various forms of qualitative information – we may not know how to value a house cost difference of $800 when we are considering houses costing $250,000, but know with high probability that $900,000 is out of our budgetary range. Thus, defining rationality in terms of quantification of distributions of variables, even under assumptions of cognitive limitations, may miss the key issue, that we must define rationality by the actions of humans facing vague information, vagueness that humans must have evolved to handle.

The importance of constraints on constraints

Christopher J. Bates[a], Chris R. Sims[b] and Robert A. Jacobs[a]

[a]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 and [b]Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180.
cjbates@ur.rochester.edu
http://www2.bcs.rochester.edu/sites/cbates/
simsc3@rpi.edu
http://www.cogsci.rpi.edu/~simsc3/contact.html
rjacobs@ur.rochester.edu
http://www2.bcs.rochester.edu/sites/jacobslab/

doi:10.1017/S0140525X19001572, e3

Abstract
The "resource-rational" approach is ambitious and worthwhile. A shortcoming of the proposed approach is that it fails to constrain what counts as a constraint. As a result, constraints used in different cognitive domains often have nothing in common. We describe an alternative framework that satisfies many of the desiderata of the resource-rational approach, but in a more disciplined manner.

A striking aspect of the Lieder and Griffiths (L&G) article is the broad generality of the proposed "resource-rational" approach (e.g., see Box 2 of the target article). It is intended to be applicable to nearly every domain in cognitive science, including perception, language, memory, attention, reasoning, and decision making. Posing a theory at this level of generality has considerable strengths, but also weaknesses. An important weakness is the lack of emphasis on identifying specific constraints that are constant across domains.

For example, L&G identify constraints operating in the domain of decision making (e.g., a person may need to minimize the amount of time required to reach a decision), which are different than the constraints operating in the domain of memory (e.g., a person may need to minimize the use of limited memory resources), which are different than the constraints operating in the domain of attention (e.g., a person may need to minimize the use of limited attentional resources), which are different than the constraints operating in the domain of reasoning (e.g., a person may need to minimize the number of variables that are reasoned about), which are different than the constraints operating in the domain of language (e.g., a person may need to minimize the amount of communication effort). Our concern is that a lack of constraints on what counts as a constraint leads to theories from different domains that have little or nothing in common. This, in turn, defeats L&G's stated purpose of providing "unifying explanations," and may lead to over-fitting and "just-so" theorizing.

It does not have to be this way. We have been pursuing a research program meeting many of the desiderata of L&G's rational resource approach but, critically, propose a common constraint across domains, namely an information-theoretic capacity constraint (e.g., Bates et al. 2019; Sims 2016; 2018; Sims et al. 2012). Our program hypothesizes that the need for efficient data compression (i.e., efficient representations) shapes biological systems in many of the same ways that it shapes engineered systems. If true, then the tool engineers use to analyze and design information-processing systems, namely rate-distortion theory, can profitably be used to understand human perception and cognition. In brief, this theory provides an optimal framework relating the limited capacity of a system to its optimal task performance.

Consider, for example, our application of rate-distortion theory to the study of visual working memory. Here, visual working memory is thought of as an information channel. When an observer encodes a visual image in visual working memory, the observer is sending a message to his or her future self. Because the message retrieved from memory will be different than the sent message – because of the memory store's limited capacity and to memory noise – the observer needs to use the retrieved message to make his or her best guess as to the sent message, thereby recalling the image. In this application, the only constraint is the capacity of visual working memory, measured as the mutual information between the sent and retrieved messages (roughly, a measure of how well the sent message can be reconstructed from the retrieved message). If visual working memory has high capacity, then the observer can recall many of the fine perceptual details of the image. In contrast, if it has low capacity, the observer will be able to recall only coarse details, such as category information (e.g., the image depicted a boy eating an apple). Although described here in an intuitive manner, a strength of rate-distortion theory is its rigorous mathematical foundation which have made it commonplace in the field of engineering.

At first glance, it may seem as if rate-distortion theory is relevant only in tasks that can be regarded as involving communication. In fact, it is relevant to any capacity-limited agent (biological or artificial) that needs to form efficient mental representations while seeking to maximize task performance. That is, it is relevant to nearly all of human perception and cognition. Admittedly, it is not always obvious how to apply this theory to many aspects of cognition. In the remainder of this commentary, we briefly describe two recent efforts to expand the application of the theory across domains.

First, we have developed a deep neural network system that approximately implements rate-distortion theory in a task-general manner (Bates & Jacobs 2019). It consists of two networks, a memory module that uses the theory to learn efficient latent representations, and a decision module that uses the memory module's latent representations to perform a task. Because of the connection between the memory module's representations and the decision module, the system learns approximately optimal representations which are both capacity-limited and task-dependent. Importantly, the system is trained "end-to-end," operating on raw perceptual input (e.g., pixels) rather than intermediate levels of abstraction, as is the case with most psychological models.

Second, we are exploring information capacity limits in human reinforcement learning. Here, the goal is to learn a behavioral policy that maximizes task performance. For example, in the game of chess, a (possible) behavioral policy might correspond to a lookup table specifying the optimal move for every board configuration. However, human learners have finite resources, and hence cannot store policies with unlimited complexity or fidelity (see also Botvinick et al. 2015). Instead, humans often learn approximate (compressed) but general policies, such as "control the center of the board." As applied to reinforcement learning, rate-distortion theory provides a precise mathematical definition of an optimal but capacity-limited policy. Capacity-limited learners necessarily acquire representations that are efficient, resulting in policies that also generalize better to novel situations (Lerch & Sims 2019).

In conclusion, consistent with the desiderata of L&G's resource-rational approach, the rate-distortion theory framework studies human cognition from an optimality perspective (similar to rational analysis), where optimal task solutions are constrained by people's cognitive architecture (i.e., capacity limits). It does so, however, in a disciplined manner that constrains what counts as a constraint. We believe this approach is necessary to achieve the ambitious and worthwhile goals set out by L&G.

Optimal, resource-rational or sub-optimal? Insights from cognitive development

Vikranth R. Bejjanki[a] and Richard N. Aslin[b]

[a]Department of Psychology and Program in Neuroscience, Hamilton College, Clinton, NY 13323 and [b]Haskins Laboratories, New Haven, CT 06511.
bejjanki@hamilton.edu
https://www.hamilton.edu/academics/our-faculty/directory/faculty-detail/bejjanki-rao
richard.aslin@yale.edu
https://haskinslabs.org/people/richard-aslin

doi:10.1017/S0140525X19001614, e4

Abstract

We agree with the authors regarding the utility of viewing cognition as resulting from an optimal use of limited resources. Here, we advocate for extending this approach to the study of cognitive development, which we feel provides particularly powerful insight into the debate between bounded optimality and true sub-optimality, precisely because young children have limited computational and cognitive resources.

We agree with Lieder and Griffiths (L&G) that when combined with reasonable assumptions about human cognitive capacities and limitations, the principle of bounded optimality provides a realistic normative standard for cognitive operations and representations. Indeed, L&G apply this standard to effectively argue that a wide range of human cognitive behaviors can be viewed as resulting from an optimal use of limited resources. We were surprised however, that they did not extend their analysis to consider human behavior during cognitive development. Specifically, we feel that cognition during early development provides particularly powerful insight into the debate between bounded optimality and true sub-optimality, precisely because young children have limited computational and cognitive resources. Moreover, there are relatively straightforward ways of estimating these resource limitations, and how they might be changing across development, rather than having to make assumptions about how limiting they might be.

Extensive research with human adults has documented that they are adept at mitigating the influence of sensory uncertainty on task performance by integrating sensory cues with learned prior information, in a Bayes-optimal fashion (Bejjanki et al. 2016; Berniker et al. 2010; Jazayeri & Shadlen 2010; Körding & Wolpert 2004; Kwon & Knill 2013; Stocker & Simoncelli 2006; Tassinari et al. 2006). Further research has shown that young children and infants are sensitive to environmental regularities, and that the ability to learn and use such regularities is involved in the development of several cognitive abilities (Fiser & Aslin 2002; Gopnik et al. 2001; Jusczyk & Aslin 1995; Kirkham et al. 2002; Kuhl & Meltzoff 1982; Neil et al. 2006; Saffran et al. 1996; Xu & Garcia 2008). However, it has also been reported that children younger than 8–12 years of age demonstrate substantial deficits in their ability to optimally mitigate the influence of sensory uncertainty by using multiple sources of information (Barutchu et al. 2008; Chambers et al. 2018; Gori et al. 2008;

Nardini et al. 2010; Nardini et al. 2008; Petrini et al. 2014). Some have suggested that the basis for this sub-optimality is a deficiency in the fundamental computational mechanism involved in combining two or more sources of information, which might take 8–12 years to fully mature. Applying the resource-rational analysis to this problem suggests an alternative possibility (as highlighted by L&G): "[children's] heuristics might already make optimal use of their cognitive resources but the computational complexity of the problem might exceed their cognitive capacities." Indeed, we have recently found (Bejjanki et al. 2019) that 6–7-year-olds are capable of integrating learned regularities with sensory information in a statistically optimal manner (that is indistinguishable from adults), provided that task complexity is reduced. Performance in tasks involving greater complexity necessitates the deployment of sophisticated top-down mechanisms (e.g., cognitive control, executive function, etc.) that typically do not reach adult-like levels until early adolescence (Best & Miller 2010; Carlson et al. 2013; Davidson et al. 2006; Luciana & Nelson 1998; Zelazo et al. 2013). Indeed, several studies have shown that young children's behavior in tasks drawing upon these mechanisms is critically moderated by task complexity. For instance, Luciana and Nelson (1998) found that while 5–7-year-olds were indistinguishable from adults when carrying out simple versions of a spatial working memory task, as task demands increased, performance in 5–7-year-olds, but not adults, deteriorated rapidly. Similarly, Davidson et al. (2006) found that while even 4-year-olds could simultaneously hold information in mind and inhibit a dominant response when rules remained constant, the ability to flexibly switch between rules was not adult-like even in 13-year-olds. Thus, children's inability to demonstrate Bayes-optimal computations in complex tasks might have less to do with their computational capacity and more to do with the immature cognitive resources that are available to them. These findings are therefore consistent with a resource-rational explanation.

More broadly, resource-rational analysis is built on the assumption that cognitive mechanisms are well-adapted to their function, and the cognitive constraints under which they operate. Although L&G briefly allude to a need to understand the process by which cognitive mechanisms are adapted to the constraints at hand via learning or evolution, they do not consider this question, or its implications for cognitive development, in any detail. For instance, their speculation that resource-rational decision mechanisms are provided by evolution or learning during development finesses the key question about how such decision mechanisms are deployed. We argue that the application of resource rational analysis would shed important new light on cognitive development. In particular, considering the bounds imposed by limited cognitive and computational resources should be, but is not currently, an important consideration in developing a normative standard for evaluating cognition across development. As illustrated above, "failures" in young children's ability to carry out sophisticated computations need not be attributed to deficits in the fundamental computational capacity available to children early in development, but rather to ancillary immaturities in general cognitive abilities. Similarly, resource rational analysis can potentially elucidate how and why young children might *outperform* adults. For instance, young children have an enhanced ability to learn languages. According to one prominent hypothesis (the *less is more* hypothesis), young children outperform adults in learning languages precisely because their resource constraints limit their ability to entertain complex hypotheses (Hudson Kam & Newport 2005; Newport 1990). Consistent with this hypothesis, Elman (1993) showed that initially resource-constrained neural networks learned grammatical structure better than unconstrained nets – resource constraints prevented the search for complex patterns, keeping networks from getting stuck in local minima. Similarly, Kersten and Earles (2001) showed that adults learned miniature artificial languages better when initially presented with only small segments of language than when they were presented immediately with the full complexity of the language. These findings, and indeed the *less is more* hypothesis, are consistent with the predictions of resource-rational analysis: given limited resources, resource rationality depends on the availability of information that can be optimally exploited by the available cognitive and computational resources.

Resource-rationality as a normative standard of human rationality

Matteo Colombo

Tilburg Center for Logic, Ethics and Philosophy of Science, Tilburg University, 5000 LE Tilburg, The Netherlands.
m.colombo@uvt.nl
https://mteocolphi.wordpress.com/

doi:10.1017/S0140525X19001596, e5

Abstract

Lieder and Griffiths introduce resource-rational analysis as a methodological device for the empirical study of the mind. But they also suggest resource-rationality serves as a normative standard to reassess the limits and scope of human rationality. Although the methodological status of resource-rational analysis is convincing, its normative status is not.

Lieder and Griffiths's *resource-rational analysis* aims to provide researchers with a methodological device to model many different kinds of cognitive phenomena in a precise way – similarly to Reinforcement Learning or Bayesian modelling (Colombo & Hartmann 2017; Colombo & Seriès 2012). Although Lieder and Griffiths explain that "resource rationality is not a fully fleshed out theory of cognition, designed as a new standard of normativity against which human judgements can be assessed" (sect. 3, para. 7), they also point out that resource-rationality can be used as "a more realistic normative standard" to revisit the debate about the scope and limits of human rationality (sect. 5.4).

Understood as a normative standard, the notion of resource-rationality encapsulated in Lieder and Griffiths's Equation 4 says that rational agents *ought to* act so as to maximise *some* sort of expected utility, taking into account the costs of computation, time pressures, and limitations in the processing of relevant information available in the environment. To contribute productively to the debate about human rationality, researchers who endorse resource-rationality as a normative standard should answer two sets of questions. First, in virtue of what does the resource-rationality standard have normative force? Why, and in what sense, is it a requirement of rationality? Second, given this standard, what does it take for an agent to make an error, to be biased or irrational?

One potentially helpful distinction to begin address these questions is between constructivist and ecological models of rationality (Colombo 2019; Smith 2008, sect. 5). Constructivist models assume that rational agents comply with general-purpose norms for successfully solving well-defined problems. Ecological models assume that rational agents are adapted to specific types of environments, where their chances of survival and their rate of reproduction are higher compared to other types of environments. Where constructivist models allow researchers to evaluate behaviour against norms, ecological models allow researchers to evaluate behaviour against organisms' objective goals of survival and reproduction.

If resource-rationality is to be understood as a constructivist normative standard, then one might try to ground its normative force in some argument similar to those typically cited in support of constructivist models like expected utility maximisation (cf., Briggs 2017; Hájek 2008, sect. 2; Colombo, Lee & Hartmann, forthcoming, sect. 3.2). There are, for example, arguments based on representation theorems, which say that if all your preferences satisfy certain "rationality" constraints, then there is some representation of you as an expected utility maximiser. There are long-run arguments, according to which if you always maximise expected utility, then, in the long run, you are likely to maximise actual utility. There are "Dutch book" arguments, which say that if your beliefs are probabilistically incoherent, there exists a set of bets you consider fair, but that guarantee your loss. And there are arguments based on accuracy considerations, which establish that if your beliefs are probabilistically incoherent, there is some probability function representing a different set of beliefs that is more accurate than your beliefs in every possible situation. There are several objections against these arguments; and in any case, it is not obvious these arguments carry over into resource-rationality.

If resource-rationality is an ecological normative standard, then the challenge is to show that, in specific types of environments, specific behavioural strategies for maximising the sort of utility encapsulated in Equation 4 promote an organism's goals of survival and reproduction. In particular, for an ecological understanding of resource-rationality to have normative teeth, researchers should show that certain strategies, which possess some epistemically good feature such as reliability, accuracy, or coherence, or which promote an organism's happiness, well-being or capabilities, approximate the resource-rationality maximum more closely than alternative strategies in many different types of realistic situations (e.g., Cooper 2001; Gintis 2009). And researchers should also show that humans employing those strategies are more likely to survive and reproduce. On pain of circularity, one cannot ground the normative force of resource-rationality by just "[p]erforming cost-benefit analyses similar to those defined in Equation 4 to determine to which extent evolution has succeeded to design resource-rational neural hardware" (sect. 5.4, para. 3).

Whether we understand the standard of resource-rationality as a constructivist norm or as an ecological goal (or both), it is not clear when violations of this standard constitute errors, or cognitive biases. There are several different norms of epistemic and practical rationality; and there probably are different kinds of goals (or cost functions) agents (or their brains) may optimise (Marblestone et al. 2016). Considering this plurality, violations of resource-rationality do not provide us with sufficient grounds for diagnosing irrationality. Furthermore, resource-rational agents "might have to rely on heuristics for choosing heuristics to approximate the prescriptions" of resource-rationality in some situations (sect. 3, para. 6). Deviating from resource-rationality cannot count as an error or a cognitive bias in those situations, unless we have a proposal about how closely behaviour should approximate the resource-rational maximum to count as (ir)rational.

One peril of using resource-rationality as a normative standard for reconsidering the debate about human rationality is that it may reiterate fruitless rationality wars. In recent years, this debate has invited "rationality wars" characterised by "rhetorical flourishes" concealing substantial empirical agreement (Samuels, Stich & Bishop 2002, 241), ambiguous use of terms such as "optimality" and "rationality" (Rahnev & Denison 2018a, 49–50), and confusion concerning the nature and methodological role of modelling approaches such as Bayesian decision theory (cf., Bowers & Davis 2012a; 2012b). To avoid this peril, researchers who are going to appeal to resource-rationality as a normative standard of rationality and contribute to the debate about human rationality should be clear on what considerations ground the normative force of resource-rationality and when deviations from this standard count as irrational errors.

Acknowledgments. I am grateful to Dominik Klein for helpful conversations on (ir)rationality, and to the Alexander von Humboldt Foundation for financial support.

Another claim for cognitive history

Henry M. Cowles[a] and Jamie Kreiner[b]

[a]University of Michigan, LSA History, 1029 Tisch Hall, Ann Arbor, MI; and
[b]University of Georgia, Athens, GA 30602.
cowles@umich.edu
https://lsa.umich.edu/history/people/faculty/henry-cowles.html
jkreiner@uga.edu
http://history.uga.edu/directory/people/jamie-kreiner

doi:10.1017/S0140525X19001547, e6

Abstract

History can help refine the resource-rational model by uncovering how cultural and cognitive forces act together to shape decision-making. Specifically, history reveals how the meanings of key terms like "problem" and "solution" shift over time. Studying choices in their cultural contexts illuminates how changing perceptions of the decision-making process affect how choices are made on the ground.

Resource-rational analysis will find friends among historians. It accords closely with an operating assumption in the field: When people seem to act against their own interests, they may be doing so in light of priorities and constraints that aren't obvious to outsiders – or even to themselves. And historians are likely to join Lieder and Griffiths in rejecting the idea of a single, ideal Rationality in favor of different "rationalities" at work in the world (d'Avray 2010).

The challenge, as Lieder and Griffiths know, is to characterize those rationalities – that is, to identify the priorities and constraints that shape specific choices. To do so, we argue, requires going beyond the fields Lieder and Griffiths invoke (psychology, economics, neuroscience, and linguistics) and looking at the contexts in which choices are actually made. We suggest history is a useful tool for doing that. History is full of examples of conscious and non-conscious factors shaping specific decisions, including instances that are baffling at first but that historical analysis helps to clarify. To take an iconic example: Why would a group of printshop apprentices murder domesticated cats in response to working conditions in old-regime France? Historians have unearthed hidden factors to explain this and other puzzles (Darnton 1984, cf. Pettit 2017).

They do so by identifying situation-specific decision mechanisms, which is just the sort of thing Lieder and Griffiths are after. Attention to context, in other words, promises to help practitioners of resource-rational analysis by altering its first step (as outlined on p. 4). Cases in cultural history suggest that identifying "a problem and its solution" is not as simple as it seems – often because the very nature of a "problem" is up for grabs. Here, we present two such cases that reveal how resource rationality is as cultural as it is cognitive.

The first case is a classic example of perplexing decision-making. In the late antique and early medieval West (roughly 350–950 CE), rich and middling donors founded hundreds of monasteries, made small gifts to local churches, built shrines to their favorite saints, paid for lighting so that certain sacred spaces would be perpetually illuminated, and arranged for prayers to be sung in honor of themselves or their families. Their gestures might seem like bad decisions. Some of the donors' heirs certainly thought so. But Christians actually had reasons to spend their lands and treasure in this way. Being generous to the poor (or to the clergy who cared for the poor) could create political capital: elites were seen as more deserving of political power when they showed concern for people with less of it. Christians also believed that this kind of spending was actually an investment that they would recoup in the afterlife. Not only were their donations not a waste of money. They were motivated by the sense that God's unique power made it possible to connect the economies of earth and heaven (Brown 2012; 2015; Kreiner 2014; Wood 2013).

What we count as resources, and what we consider rational uses of those resources, will change over time. But even more fundamentally, this example of Christian expenditure points to the fact that the moments when we ask ourselves, "Should we spend these resources, and how?" are historically determined. We are cued by our culture to diagnose dilemmas. Giving land to a church had not always been perceived as a solution to a problem. The "problem" itself – the recognition that there was a choice to confront about profit, with different outcomes to consider – had not always existed. It took centuries of preaching, arguing, and storytelling to get to the point where an elite person could be expected to see the choice to donate (or not) as a possible response to a self-evident challenge.

If the lesson of early Christian charity is that new behaviors can eclipse the rationality of older ones, the second case shows that the meanings of problems and solutions can change under our feet. This case centers on the idea of "conspicuous consumption" made famous by Thorstein Veblen's *Theory of the Leisure Class* (Veblen 1899). Drawing on evolutionary biology, Veblen argued that seemingly simple consumer decisions were in fact elaborate performances meant to reveal our adaptive fitness (Raymer 2013). Buying a fancy watch, for instance, wasn't (just) a solution to the problem of telling time. It was also, according to Veblen, a solution to a much deeper, unrecognized problem: signaling strength to potential competitors and mates.

This view of decision-making was controversial, to say the least. Criticisms came from all political and intellectual sides (Tilman 1991), and even Veblen's fans deemed him (as one biographer put it) "the bard of savagery" (Diggins 1978). But the point stuck: economic choices – and indeed, choices in general – often solve problems we are not even aware of. On the surface, consumption seemed like a straightforward negotiation between supply and demand; the resource-rationality of a purchase appeared self-evident. The effect of Veblen's argument was to complicate how economic decisions were defined as problems and solutions. Even if his theory was wrong in the particulars, its historical emergence is a reminder that resource-rational analysis depends on ideas (consumer preference, game theory) with their own histories.

We present these case studies not to deflate the value of resource-rational analysis but rather to enrich it. History can help capture the stranger aspects of human cognition, by drawing attention to the ways that people have come to count (or not count) certain things as problems, choices, and solutions in the first place. And if resource-rational analysis achieves wider recognition, an attention to how "problems" and "solutions" are defined could have an intriguing secondary effect. In the hands of actual people making actual decisions, it could feed into the very processes that Lieder and Griffiths document. That is, resource-rational analysis could be an engine, not a camera (MacKenzie 2006), altering how people understand the problems they face and the solutions available to them.

Computational limits don't fully explain human cognitive limitations

Ernest S. Davis[a] and Gary F. Marcus[b]

[a]Department of Computer Science, New York University, New York, NY 10012 and [b]Department of Psychology, New York University, New York, NY 10003.
davise@cs.nyu.edu
https://cs.nyu.edu/faculty/davise/
gary.marcus@nyu.edu
http://garymarcus.com/

doi:10.1017/S0140525X19001651, e7

Abstract

The project of justifying all the limits and failings of human cognition as inevitable consequences of strategies that are actually "optimal" relative to the limits on computational resources available may have some value, but it is far from a complete explanation. It is inconsistent with both common observation and a large body of experimentation, and it is of limited use in explaining human cognition.

Advocates of the oversold view that human cognition is "optimal" are in the midst of a strategic retreat. If it no longer looks like human cognition is optimal, might it be "bounded optimal," optimal relative to inherent limits on information, computational resources, cognitive capacity, or neuronal architecture? In the limit, the claim is meaningless: whatever the brain does is constrained by whatever the architecture is. But even a less vague version, centered around limits on memory, information, or computational power as an explanation for cognitive flaws, yields little traction.

Consider the laziness doctrine. The flagship class of theories in Lieder and Griffiths' paper is a large body of work on decision making. Many departures from normatively optimal decision-making can be explained on the supposition that finding the optimal action requires more mental effort than seems worthwhile. But a serious version of the bounded rationality view must presume that the tradeoff between effort and decision-making is made *optimally*. Any reader who thinks honestly about their own decision making will probably recognize occasions on which they have occurred incur large, foreseeable costs, because they were too impatient; a bland claim that people manage informational trades optimally is at odd with everyday reality.

Toward the end of the paper, Lieder and Griffiths raise the issue of "everyday observations of seemingly irrational beliefs and behaviors," but then give the comforting explanation that those must be beliefs of no adaptive significance, like whether the world is flat, so the human is wise not to spend any cognitive effort on them. But that does not explain behaviors that foolishly risk one's life, such as drunk or careless driving, or the hundreds of people who have died taking selfies, misjudging fatal risks in the pursuit of a few more followers on Instagram. The trouble is, Lieder and Griffith's approach sounds nice but predicts very little of the texture of actual human decision making.

Lieder and Griffiths also cite numerous studies claiming that human memory is bounded-optimal in some respects. In fact, as one of us (Marcus 2008) has argued at length, memory is a very clear case of a suboptimal system. Memory lapses of salient and important realities are notoriously common and often costly; parachutists have been known to forget to pull their ripcords, and airline pilots have checklists precisely because human memory cannot be trusted in life or death situations. Meanwhile, the existence of mnemonic tricks like the method of the loci show that the mental limitations of ordinary humans are not *inevitable* limitations of a neuronal architecture, because, with training and practice, ordinary limits can be substantially overcome. That said, our default memory systems just aren't that good. And the notion of bounded optimality casts virtually no light on what is and is not easy. It tells us little about why, say, we can recognize hundreds of faces of people in high school that we haven't seen for decades, yet fail to remember a 10-digit passport number or where we parked three hours earlier in a shopping mall parking lot.

An addiction to the presumption that all must be optimal, if only the right resource-limitation can be found, such that erroneous behavior can be executed, leads to all kinds of weird reasoning. Lieder and Griffiths write, for example, that "Rational models … have provided surprisingly good explanation of cognitive biases …. Includ[ing] the confirmation bias," and cite Oaksford and Chater (1994) and Austerweil and Griffiths (2011) in support. To get there, Austerweil and Griffiths narrowly define the confirmation bias as "the tendency to test outcomes that are predicted by our current theory" and demonstrate that that is an optimal strategy if one is testing deterministic causal laws; Oaksford and Chater's analysis is similar But the usual meaning of confirmation bias is much broader, for example (Plous 1993): "the tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses." We all know, and it has often been systematically demonstrated, that someone who believes that the moon landings were faked (say) is likely to attend to, emphasize, and remember any and all evidence that supports this theory and to ignore, discount, and forget all contrary evidence; and no kind of argumentation will convince us that this is rational. These two studies are hardly enough to address the broader sense. In fact the target paper by Lieder and Griffiths is *itself* an instance of confirmation bias in this broader sense: it is an enumeration of cases that might possibly be construed of as limitation-induced cognitive bias, without anything like careful analysis of the scope of other cases that might fall outside that scope.

Uncovering cognitive constraints is the bottleneck in resource-rational analysis

Cvetomir Dimov

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.
cdimov@andrew.cmu.edu

doi:10.1017/S0140525X19001675, e8

Abstract

A major constraint in resource-rational analysis is cognitive resources. Yet, uncovering the nature of individual components of the human mind has progressed slowly, because even the simplest behavior is a function of most (if not all) of the mind. Accelerating our understanding of the mind's structure requires more efforts in developing cognitive architectures.

Rational analysis has found multiple applications in generating behavioral predictions based on task structure (e.g., Oaksford & Chater 1994; Sims et al. 2013). Yet, the initial two uses of this method have demonstrated its potential for more. First, the rational analysis of memory (Anderson & Milson 1989; Anderson & Schooler 1991) did not only predict behavior, but, in fact, developed a theory of that cognitive capacity. This theory has been successfully applied to many memory (e.g., Anderson et al. 1998; Schneider & Anderson 2012) and decision-making tasks (e.g., Dimov & Link 2017; Fechner et al. 2016). Second, the rational analysis of categorization (Anderson 1991) also achieved more than predict behavior: Keeping considerations of cognitive plausibility in mind, an *algorithm* was developed that sequentially assigns stimuli to categories. To summarize, in addition to predicting behavior, rational analysis can be used to develop theories of cognitive capacities given environmental constraints and, second, it also penetrates the algorithmic level when necessary to explain experimental data.

With their resource-rational analysis, Lieder and Griffiths extend rational analysis by including cognitive constraints into the optimization function. The viability and explanatory power of resource-rational analysis is well supported by many successful applications. However, many of these applications have one thing in common: They are relatively independent of the underlying cognitive capacities. Thus, one major constraint of resource-rational analysis – computational resources – is often avoided, in many cases reducing the approach to that of finding the optimal algorithm under task constraints – a procedure strikingly similar to rational analysis. The reason, also acknowledged by the authors themselves, is that measuring cognitive constraints has progressed slowly.

Behind this slow progress lies a fundamental difficulty in reverse-engineering the nature of the cognitive system's components. As argued by Newell (1990), each psychological experiment produces output that is the joint product of several cognitive processes. Consequently, data collected to advance our understanding of one cognitive process is marred with unexplained variance from several others, which renders determining the exact structure of the process under investigation problematic. To address this issue, Newell proposed iteratively developing and refining a model of the entire cognitive system – a *unified theory of the mind* – that provides a unified account of an ever-increasing number of psychological tasks. By jointly carving away unexplained variance from all components of the mind, such a theory would enable each subsequent experiment to ask more specific questions about the psychological process it investigates.

Newell's behest was followed by several *cognitive architectures*, the most developed among which is likely ACT-R (Anderson 2007). This architecture has incorporated the rational analysis of memory into the earlier ACT* (Anderson 1983), added the perceptual and motor processes, meticulously developed for the *EPIC* cognitive architecture (Meyer & Kieras 1997a; 1997b), and linked its components to regions in the brain (Anderson et al. 2016; Borst & Anderson 2017). Currently, it is able to account for behavior in hundreds of tasks in various fields, which include language learning and comprehension (Budiu & Anderson 2004; Taatgen & Anderson 2002), decision making (Marewski & Schooler 2011), driving (Salvucci & Taatgen 2010), and many others (see http://act-r.psy.cmu.edu/publication/ for a list of publications categorized by field). This likely makes ACT-R the best source of cognitive constraints for resource-rational analysis.

Whether ACT-R is the cognitive theory of choice or not, Newell's arguments remain valid today: addressing the *identifiability problem* (Anderson 1990), the *irrelevant specification problem* (Newell 1990), or the problem of *amortization of theoretical constructs* (Newell 1990) is likely to be most successful with a unified theory of the mind that progressively incorporates multiple constraints from experiments, evolutionary arguments, and functional considerations. In my opinion, we should be devoting more efforts to develop such theories to accelerate our understanding of the mind as even the leader, ACT-R, despite its many successes, is still far from complete: It lacks theories of some fundamental components of the mind, such as emotions and tactile and other sensations, whereas many of the currently included components will likely be subjected to multiple refinements and extensions as this architecture is put to the test in new tasks.

Advancing a unified theory of the mind will naturally benefit approaches such as resource-relational analysis. Moreover, I believe that this approach might play a role in unveiling the structure of the mind similar to the role rational analysis played in developing a theory of memory. Specifically, if we maintain the assumption of optimality, we can ask under what cognitive and task constraints the empirically observed algorithms would be optimal, which could allow us to narrow down the plausible region in the space of possible computational resources. Such synergies between optimization approaches and cognitive architectures coupled with growing efforts in developing the latter will likely lead to considerable advancements in our understanding of human cognition and behavior.

Resource-rationality beyond individual minds: the case of interactive language use

Mark Dingemanse[a,b,c]

[a]Centre for Language Studies, Radboud University, 6525 HT Nijmegen, Netherlands; [b]Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 HR Nijmegen, Netherlands and [c]Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, Netherlands.
m.dingemanse@let.ru.nl
https://www.ru.nl/english/people/dingemanse-m/

doi:10.1017/S0140525X19001638, e9

Abstract

Resource-rational approaches offer much promise for understanding human cognition, especially if they can reach beyond the confines of individual minds. Language allows people to transcend individual resource limitations by augmenting computation and enabling distributed cognition. Interactive language use, an environment where social rational agents routinely deal with resource constraints together, offers a natural laboratory to test resource-rationality in the wild.

The target article sketches the promise of combining rational principles and cognitive constraints to understand human cognition, and singles out linguistics as one domain for work along those lines. While it touches on aspects of language rooted in individual cognition like the principle of least effort (Lestrade 2017; Zipf 1949), I want to probe the limits of the resource-rational framework by looking beyond individual minds to interactive language use, the primary ecology of human cognition (Böckler et al. 2010; Waldron & Cegala 1992). Here, under the relentless pressures of rapid-fire turn-taking (Levinson 2016) and always-on inferential processes (Enfield 2013; Goffman 1967), language provides a window onto how social rational agents deal with resource limitations in a noisy and uncertain environment.

Human language provides ample evidence of adaptation to capacity limits in social interaction (Roberts & Levinson 2017). Articulation, relatively slow compared to processes of formulation and interpretation, forms a significant bottleneck in human communication that we can bypass thanks to pragmatic inference (Levinson 2000): any content that can be left to inference need not be explicitly articulated. This puts a premium on computable and efficient heuristics for formulation and interpretation (Frank & Goodman 2012; Van Rooij et al. 2011). But as Lieder and Griffiths argue, people cope with computational complexity through heuristics as well as through habits. One way to think of language is as offering a culturally evolved store of habits – routinely deployable resources – that help outsource computation and streamline coordination (Clark 1998; Kempson et al. 2016).

A resource-rational approach may be especially promising for understanding the ubiquity of delay markers, continuers, and repair strategies, which easily occur in up to one in five utterances (Enfield 2017; Fox Tree 1995). Whereas classic linguistic work has assumed such items are grammatically irrelevant (Chomsky 1965) or at most symptoms of trouble (Levelt 1989), resource-rationality makes it possible to account for them as optimally adaptive interactional tools (Dingemanse 2017): cognitive crutches that help optimize complex rational communication under resource limitations. For instance, delay markers like "um" help word recognition by alerting the recipient that an upcoming word might need more attention (Fox Tree 2001), and repair initiators like "huh?" or "who?" allow us to gracefully recover from impending communicative trouble, something that happens, on average, at least every 84 s in conversation (Dingemanse et al. 2015). With interactional tools available at every turn to review, revise, and recalibrate understanding, the dynamics of human cognition in interaction diverges radically from the one-shot models assumed in many current theories.

As a consequence, interactive language use calls into question the exclusive focus of rational analysis on individual minds. Are resource-rational approaches limited to individual cognition or could they extend to socially distributed cognition? By enabling the redistribution of attentional, cognitive, and embodied resources (Clark 2006; Hutchins 1995), interactive language use alleviates individual-bound capacity limits and can optimize performance beyond the bounds of idealized one-shot communication: an interactively scaffolded form of cognitive offloading (Risko & Gilbert 2016). The sheer frequency of the interactional tools mentioned above shows how much communication relies on this form of scaffolding (Fusaroli et al. 2017). This radically increases the error-tolerance and flexibility of cognition in interaction. It also creates opportunities to study the workings of resource-rationality in the relatively controlled environment of well-understood sequential patterns of interaction.

Communicating under noise and uncertainty requires constant cost-benefit analyses of formulating a response versus issuing a request for repair, factoring in the relative costs of different repair formats and their possible downstream consequences, all under severe time pressure and with limited cognitive resources. A systematic comparison of repair across languages and cultures shows that people everywhere deploy the repair system in efficient ways that minimize cost for the dyad as a social unit, rather than just for themselves as individual-based rational approaches might suggest (Dingemanse et al. 2015): an optimal use of *distributed* cognitive resources. A similar interactive, distributed perspective is required to make sense of information-theoretical results about word meanings and ambiguity (Piantadosi et al. 2012): we can cope with ambiguity in communication only to the extent that one mind picks up the slack where the other leaves off. This means that resource-rational analysis of human cognition will need to deal not just with individual minds, but with interacting minds operating in an environment of culturally evolved metacognitive resources.

Recent work in cognitive science and cultural evolution is revisiting the Vygotskyan insight that human cognition is greatly amplified by culturally evolved pieces of cognitive equipment (Bender & Beller 2014; Clark 2006; Heyes 2018). At the same time, neuroscience is increasingly concerned with understanding brain and language in the context of social interaction (Hirsch et al. 2018; Konvalinka & Roepstorff 2012; Schilbach et al. 2013). One thing that unites these approaches is their attention to how the picture of cognitive demands and resources may change radically as a result of interactionally scaffolded, socially augmented cognition. Lieder and Griffiths do not discuss cultural evolution and social interaction as part of the environment in which heuristics and habits can be honed to become optimally adaptive, and it is unclear whether they intend resource-rational analysis to include the kinds of interactional resources discussed here: material symbols of metacognition that augment and distribute our cognitive processes. Perhaps this is the next frontier.

In sum, I applaud the call for new ways to connect psychological theory and the cognitive sciences, and would like to put forward interactive language use as a challenging yet promising domain for resource-rational approaches. As the primary ecology of human cognition, social interaction provides a rich natural laboratory for probing the leverage and limits of resource-rational analysis. Future work in this vein might focus not just on how structural aspects of language adapt to the resource limitations of individual minds, but also on how every language offers its own compendium of culturally evolved ways by which people transcend individual resource limitations and benefit from distributed cognition.

Acknowledgment. I am grateful for thought-provoking conversations with Mark Blokpoel and Iris van Rooij, and for funding from NWO grant no. 016.Vidi.185.205.

Generalization of the resource-rationality principle to neural control of goal-directed movements

Natalia Dounskaia and Yury P. Shimansky

College of Health Solutions, Arizona State University, Phoenix, AZ 85004.
natalia.dounskaia@asu.edu
yury.shimansky@asu.edu
https://chs.asu.edu/natalia-dounskaia

doi:10.1017/S0140525X19001559, e10

Abstract

We review evidence that the resource-rationality principle generalizes to human movement control. Optimization of the use of limited neurocomputational resources is described by the inclusion of the "neurocomputational cost" of sensory information processing and decision making in the optimality criterion of movement control. A resulting tendency to decrease this cost can account for various phenomena observed during goal-directed movements.

Lieder and Griffiths demonstrate a capacity of the principle of optimal use of limited computational resources (resource-rationality principle) to account for a wide variety of observations in multiple disciplines, including psychology, neuroscience, linguistics, and economics. However, they have overlooked recent developments in the field of neural control of human goal-directed movements where the explanatory power of that principle has been demonstrated. We briefly review those developments below and show how several different pieces of evidence support the resource-rationality principle.

The first step toward making a connection between cognitive resources and characteristics of human motor performance was made yet by Fitts (1954). He demonstrated that if experiment participants are asked to move to a target, the movement time is linearly proportional to the index of task difficulty that Fitts described as the amount of information needed to be processed to achieve required precision. When the distance to the target is fixed, difficulty in task is determined by the size of the target, and the smaller the size the longer the movement time, which is known as speed-accuracy tradeoff. This relationship shows that the neural system controlling movements can tailor the precision of information processing to the required precision of motor task performance and movement speed to precision demands. The phenomenon of speed-accuracy tradeoff has a simple interpretation. If the movement is too fast, there may not be enough time for accumulating the minimal amount of information required for sufficiently precise state estimation and decision making.

Later studies have corrected the concept formulated by Fitts and showed that the movement control system uses a two-phase strategy. In particular, Shimansky and Rand (2013) demonstrated that during the initial phase, the use of neural resources for processing sensory information is minimized, while the control system relies on the internal model of the controlled object's dynamics. Speed-accuracy tradeoff is violated in that phase. The final phase is performed with speed-accuracy tradeoff, with the precision of sensory information processing being determined by the required movement precision at the target.

To account for their findings, Shimansky and Rand (2013) suggested that, since neural computations involved in information processing are costly, the cost of the "neural effort" should be minimized whenever possible during performance of motor tasks. Using the optimality approach, they formally introduced the "neurocomputational" cost (the cost of neural effort, or the cost of cognitive resources in terms of Lieder and Griffiths) as a vital component of the criterion (called "utility function" by Lieder and Griffiths) determining movement control optimality. Thus, the concept of two-phase strategy can be viewed as a generalization of the resource-rationality principle to reaching movement control.

The notion of the neurocomputational cost was further used to account for a hierarchical organization of control of the limb's joints that is typically observed during human movements (Dounskaia and Shimansky 2016). Namely, different joints of the limb (e.g., the shoulder and elbow) usually play different roles in movement production. One ("leading") joint is rotated actively by the muscles spanning the joint, while the other joint "trails" due to passive factors, including gravitational torque and "interaction torque" caused by motion of the leading joint. This "trailing joint control pattern" is analogous to cracking a whip by swinging its handle, although the trailing joint musculature can interfere and adjust motion of this joint to task requirements.

Multiple studies have demonstrated the trailing joint control pattern during various types of arm movements (for review, see Dounskaia 2005; 2010). As discussed by Dounskaia and Shimansky (2016), this pattern is a result of movement optimization, which is apparent from a tendency to maximally exploit passive torques for rotation of the trailing joint and from an observation that the contribution of passive torques to control of the trailing joint increases with development of skill. Dounskaia and Shimansky (2016) used the information theory to show that the trailing pattern decreases the neurocomputational cost by reducing the amount of information that needs to be processed for joint coordination. Indeed, active control and coordination of all joints requires estimation of joint positions and development of corrective control commands at each moment of time. The trailing control pattern allows delegation of joint coordination mainly to passive torques and spinal reflexes to reduce the need for expensive neurocomputational processing of external sensory and proprioceptive information at the cerebral cortical level.

The tendency to reduce the neurocomputational cost has a strong potential to account for many other motor control phenomena. For example, causes for differential stability of various multi-limb coordination patterns, including bimanual movements, remain an object of debates (Swinnen 2002). A comparison of different coordination patterns in terms of cognitive resources required for state estimation and decision making during generation of corrective control commands to each limb is a promising approach to account for experimentally observed differences in pattern stability. Theories that suggest simplification of control, for example, through the use of muscle synergies and motor primitives, and through reducing movement variability relevant for the task and ignoring irrelevant variability (Bruton and O'Dwyer 2018; Gisztér 2015; Scholz and Schoner 1999) implicitly represent the tendency to minimize this cost.

Finally, an application of the principle of the neurocomputational cost minimization to human movement control suggests that a learning process contributes to emergence of strategies that minimize the use of cognitive resources. This hypothesis is supported, for example, by an observation that more skillful movement performance is associated with the use of a more pronounced trailing joint control pattern, that is, the more intensive use of passive torques and spinal neural circuitries for production of training joint motion (Dounskaia and Shimansky 2016).

In conclusion, the extension of the resource-rationality principle to the field of human movement control described here increases the generality of this principle. This generalization will help to advance the principle of resource-rationality in both cognitive and motor control research fields.

Holistic resource-rational analysis

Julia Haas[a] and Colin Klein[b]

[a]School of Philosophy, Australian National University, Canberra 0200, ACT, Australia and [b]School of Philosophy, Australian National University, Canberra 0200, ACT, Australia.
Julia.haas@anu.edu.au
Colin.klein@anu.edu.au
http://www.juliashaas.com
http://www.colinklein.org

doi:10.1017/S0140525X19001493, e11

Abstract

We argue that Lieder and Griffiths' method for analyzing rational process models cannot capture an important constraint on resource allocation, which is competition between different processes for shared resources (Klein 2018, *Biology and Philosophy* **33**:36). We suggest that *holistic* interactions between processes on at least three different timescales – episodic, developmental, and evolutionary – must be taken into account by a complete resource-bounded explanation.

We applaud Lieder and Griffiths' focus on resource-rational explanations. We also think that it is incomplete. Their proposed top-down method for analyzing rational process models is *atomistic*. That is, it starts with an individual process and determines the optimal tradeoffs between resource usage and other desiderata. Done well, this constrains the empirical search space to a class of algorithms or even a token algorithm. These analyses are valuable. Yet atomistic analyses cannot capture an important constraint on resource allocation, which is competition between different processes for shared resources (Klein 2018). We suggest that *holistic* interactions between processes on at least three different timescales – episodic, developmental, and evolutionary – must be taken into account by a complete resource-bounded explanation.

First, consider interactions between processes on the timescale of task performance. We are capable multitaskers. Conversation while sight-reading is demanding but possible. But, some tasks that are trivial to do on their own interfere with one another when performed simultaneously. For example, it is difficult to simultaneously remember a three-digit number and do mental arithmetic. Connectionist models indicate that these limitations emerge from the *multiplexed* structure of control representations. Multiplexing refers to the strategy of using the same control representation across multiple task-domains, resulting in a limit in the number of tasks that can be performed at the same time (Botvinick & Cohen 2014; Cohen et al. 1990). Importantly, the resource-bounds which drive the explanation of task conflict cannot be derived from considering either task in isolation. Only a holistic resource-rational analysis can show the tradeoffs between processes which compete for the same computational resources.

Analogous arguments apply at the learning and evolutionary timescales. Optimization of control processes occurs through learning over time. For example, native bilingual speakers use overlapping brain circuitry to support comprehension and production in both languages, and different contexts place different demands on these shared resources. Further, Green and Abutalebi (2013) demonstrate how control representations for language switching are parameterized over developmental time in a context-sensitive way. A child develops its capacity to switch between Spanish at school and English at home. These control processes will have to adapt as the child becomes more proficient in each language, and as they encounter new contexts with new demands. Crucially, this optimization cannot be performed for each developmental stage and context independently: efficient allocation of neural and computational resources must take into account inter-process interactions.

Higher-order optimization processes also occur on evolutionary timescales. Evolution puts harsh demands on possible forms. Evolution often satisfices rather than optimizes (Simon 1996), and what *can* evolve often depends strongly on what already *has* evolved (Brown 2013). This is a point which is made in the context of the re-use of information in gene regulatory networks (Calcott 2014) and the re-use and overlap of neural implementations (Anderson 2010). We suggest that it is equally well applied to the computational and algorithmic domains with which Lieder and Griffiths are concerned. For example, the problem of mobilizing cognitive control is thought to be solved by using reward-based learning algorithms (Botvinick & Braver 2015). Given the phylogenetic breadth of reward learning, this may represent the re-use of an evolutionarily older algorithm. The search for particular first-order algorithms thus cannot be undertaken in isolation, but should be constrained by evolutionary considerations.

Science must start somewhere, and we think that the atomistic method proposed by Lieder and Griffiths is a useful way to begin empirical investigation. Yet, analyses which focus only on a single task must necessarily leave free parameters in order to incorporate potential resource competition. Thus, the pitfall of underdetermination, for which they rightly criticize others, can return for atomic resource explanations in a modified form. We believe that Lieder and Griffiths do have resources to tackle this problem, some of which are hinted at in their target article. To be fully satisfying, holistic attention to inter-process coordination will be especially important if the theory is to avoid vacuity.

Heuristics and the naturalistic fallacy

Christopher J. Kalbach

Philosophy Department, Florida State University, Tallahassee, FL 32306-1500.
CKalbach@FSU.edu
www.ChrisKalbach.com

doi:10.1017/S0140525X19001523, e12

Abstract

Lieder and Griffith's account of resource-rationality relies heavily on a notion of teleology. In this commentary, I criticize their teleocentric view as being incompatible with evolutionary theory, in which they aim to ground their analysis. As such, to save their view, I argue that they must jettison the notion of teleology, and their teleologically laden conclusions.

Falk Lieder and Thomas L. Griffiths lay out a dynamic and relevant analysis of heuristics and the rationality of using these resource-maximizing cognitive devices. However, throughout this informative discussion, a specter lurks: *telos*. The authors describe the mind as having an "ingenious design" and point out how well adapted it is for operating in our natural environment (sect. 1, para. 3). However, to try to understand rationality in accordance with this "ingenious design" is to conflate a descriptive "is" with a prescriptive "ought." This use of the naturalistic fallacy – which, for the present purposes, I will use interchangeably with the Is/Ought Fallacy – by taking an "is" to be an "ought" is dangerous in a theory of the mind, as we are likely to make inferences not justified by the "is" (Hume 1739–40/2000; Moore 1903). If evolution is the dominant background theory for psychology, which is a process by which random genetic mutation provides a procreative advantage thereby passing down the advantageous genes (Hall 2007; Hartwell et al. 2011; Herron & Freeman 2013), then we cannot make normative claims about rationality as operating in accordance with designed capacities. There is no prescriptive "ought"; there is merely the descriptive "is."

Once the prescriptive and the descriptive have been confused, it should not come as a surprise that, "resource-rationality also allows us to answer teleological questions about the nature of the mind" (sect. 7, para. 1). The notion of teleology here is one of purpose and design, neither of which is supplied by an evolutionary framework. The standard move would be to back off of the claim and say that it is just the appearance of design, sometimes called teleonomy (Pittendrigh 1958), or say that we use the metaphor of design as a simplifying assumption – a heuristic – for explanations (Ruse 2017). However, this is clearly not what the authors have in mind as they list teleological questions like, "what the purpose of goals is" (sect. 7, para. 1). According to evolutionary theory, the capacity to set goals was not adapted to perform an action or to realize a goal. The evolution-theoretic answer to these questions is simply this: because these developed capacities were useful in the environment, they provided a comparative advantage which allowed those with the capacities to reproduce and passed down the advantageous genes (Hall 2007; Hartwell et al. 2011; Herron & Freeman 2013). To use the analogy of the famed Darwinist, Michael Ruse, "No one would ask about the purpose of the meteorite that smashed into the earth some sixty-six million years ago… It just happened. There was no purpose to it" (Ruse 2017).

Now, in light of the evolutionary answer, how can we make the claim: "If we were able to derive what people's cognitive capacities should be, this would provide a very principled starting point for resource-rational analysis" (sect. 5.4, para. 3). It is difficult to make sense of what these capacities should be, because the "should be" relies on a purpose or design. As such, unless the authors are moving to a theistic (or similar) framework (Nagel 2012; Robinson 2007), the use of purpose must be jettisoned. However, the claim that we can derive what these capacities should be serves as the basis for redefining "cognitive bias" in terms of the violation of "resource-rationality" (sect. 5.4, para. 2). This results in the dubious claim that when the belief is not immediately important, "having questionable beliefs about X is not inconsistent with being (approximately) resource-rational" (sect. 6, para. 3). Now, the authors struggle to adhere to this new standard as they say, "cognitive scientists must have strong inductive biases to infer cognitive mechanisms from limited data" (sect. 1, para. 2). These cognitive scientists seem to be resource-rational given the limited data, and the authors just redefined these biases as the violation of resource-rationality.

If cognitive biases only apply to those not acting in accordance with resource-rationality (sect. 5.4, para. 2), and having questionable beliefs can be resource-rational (sect. 6, para. 3), then we find some absurd conclusions. On this view, philosophers, cognitive scientists, physicists, and the like who do not use heuristics to exploit these evolved capacities are cognitively biased, but philosophers who use heuristics to bypass difficult problems are not? Surely, the inductive bias pointed out by the authors is just exploiting an evolved capacity. As such it should not be considered a bias at all as "resource rational analysis will almost invariably favor a simple heuristic over optimization… because it penalizes… the cost of mental effort" (sect. 3, para. 5). On this account, Timon's belief that the stars are "fireflies that, uh … got stuck up on that big bluish-black thing" (Allers & Minkoff 1994) seems to be managing cognitive resources brilliantly. After all, Timon's belief is not immediately important and requires little cognitive effort, so this questionable belief can be resource-rational. Now, when compared to Galileo's dedication of reason to understanding the distant truths, Galileo is clearly not resource-rational (Galileo 1632/2001). Surely, this is the wrong result.

The problems that I traced out here all develop from the confusion of an "ought" and what is really an "is." Rationality is normative, and it carries with it prescriptive force. This rationality should be judged by the choice of the tool that is most likely to achieve the most accurate – best – outcome. On this standard view, there are many cases in which using heuristics is rational, for instance trying to catch a baseball (Gigerenzer 2010). This is because trying to calculate the parabolic curves is computationally intractable. Why is it rational in this case? Because, in this case, they provide a comparative advantage. However, heuristics are known to fail, and to heavily favor them will not achieve the most accurate outcome. To redefine cognitive bias and rationality to fit the design of these capacities is to take the descriptive fact that we have these capacities and derive the normative claim about rationality. This is the first fruit of an illicit prescription based on the naturalistic fallacy.

The biology of emotion is missing

Katherine Peil Kauffman[a,b]

[a]EFS International, Kirkland, WA 98033 and [b]Institute of Systems Biology, Seattle, WA 98109.
ktpeil@outlook.com
www.emotionalsentience.com

doi:10.1017/S0140525X19001511, e13

Abstract

Although augmenting rational models with cognitive constraints is long overdue, the emotional system – our innately *evaluative "affective" constraints* – is missing from the model. Factoring in the informational nature of emotional perception, its explicit *self-regulatory* functional logic, and the predictable pitfalls of its hardwired behavioral responses (including a maladaptive form of "identity management") can offer dramatic enhancements.

Although the resource-rationality approach is an excellent step in the right direction, in terms of how real people actually operate, the theoretical framework remains deeply inadequate. It may work well when gathering information for decision-making (toward maximizing utility, subject to budget constraints), but what about the more puzzling phenomenon of "vaccine hesitancy"? Parents refusing to vaccinate their children despite the safety, efficacy, and broad availability of vaccines (World Health Organization 2019)? Even stubbornly refusing to accept scientific evidence? What utility function are they maximizing in downright refusing information?

Is this more evidence of a cognitive architecture evolutionarily honed for quick and dirty intuitive judgments (Gilovich et al. 2002)? Hardwired constraints perhaps "mismatched" to contemporary environments (Tooby & Cosmides 2000) – outdated, error prone, dysfunctional? Or might something more biologically meaningful be happening here?

A foundational problem is that *the emotional system* – our innately *affective* computational capacities – is missing from the model. When the pleasurable and painful feeling categories are considered, it becomes clear that an *evaluative information gathering process* happens first, influencing, coloring, filtering subsequent cognitive perceptions and rational deliberations. Discussions of the "affect heuristic" (Slovic et al. 2002) have begun charting this territory, but the *informational nature* of emotional experience predates the emergence of neural structures ("cognition" proper) and carries a much deeper functional significance.

As I have argued elsewhere (Peil 2012; 2014), the *chemistry of emotion* evolved very early on in our single-celled ancestors, central to both *sensori-motor control* and *adaptive immunity* – once a singular Pangea-like function best described as "self-regulation." Its self-regulatory informational dimension is born of the very self-organizing dynamics (Kauffman 1993; Walleczek 2006), self-maintaining agentic constraints (Mossio & Moreno 2015), and entropy delaying principles that characterize life (Davies 2019), those still undergirding the genetic, epigenetic, and immune regulatory networks that define and maintain multicellular organisms. Specifically, instantiated on transmembrane receptors, this chemistry delivers *a three-step cybernetic control loop* (a common engineering control principle in machines from thermostats to guided missiles and intelligent robots). It works like this: (1) An ongoing: *comparison* is made between the "self" and its "not-self" environment; (2) a *signal* occurs when imbalances are detected, which (3) triggers a *self-correcting behavior* that rebalances the system.

More generally, this self-regulatory chemistry still drives bacterial "info-taxis" (Bray 2009), suggesting that *emotion was the first sensory system* to emerge on the evolutionary stage (Peil 2014), with both the signal and its coupled corrective response (steps 2 and 3 of the control loop) experienced subjectively as *hedonic qualia*. No matter how it evolved, emotional sentience provided tremendous selective advantage, arming even the simplest organisms with the ability to *sense* and *evaluate* environmental affordances (Gibson 1982) as "good for me" or "bad for me" and *respond correctively* with approach or avoidant behaviors – even leaving behind *memory* traces for *anticipatory* responses. Indeed, the melding of binary feelings with bodily reactions undergirds all learning systems, Pavlov's (1927) "unconditioned" stimulus–response pair, the innately rewarding and punitive evaluative categories upon which more cognitively complex judgments, attitudes, motives, and habits are forged. Perhaps most importantly, this simple regulatory control chemistry instantiates the first crude sort of *mind*, an "enactive" or "5E" mind (Peil 2017; Rowlands 2010; Varela et al. 1991), one fundamentally *embodied* in living material, cyclically *enacted* in real time, inseparably *embedded* in its local environment, *extended* through learning and niche construction, and *evaluative* given the central role of hedonic qualia. Such an emotionally in-formed mind affords living creatures direct participation in evolution, with the later neural enhancements adding more specific informational dimensions (via the need-oriented appraisal themes of basic and complex feelings) (Peil 2012).

It is difficult to overstate the theoretical implications of the self-regulatory function of emotion. In terms of somatic (pre-neural) identity, this chemistry instantiates the "proto-self" (Damasio 1999), the self/not-self distinction of the immune system (Pert 1998), is a central mediator of epigenetic ("not-yet-self") development (Radley et al. 2011) and a likely suspect in placebo and nocebo effects (Peil 2014). Misunderstanding the self-regulatory nature of emotional experience (and its hardwired behavioral safeguards) undergirds many of our problematic decision-making heuristics, "self-serving" biases, ego defenses, and unconscious behaviors. Ignorance of its *informational dimension* predicts a dysfunctional – largely pain-driven – pattern of *identity management* that fuels defense of narrow identity boundaries, competitive interpersonal conflict, political polarization, and religious fundamentalism. In terms of "anti-vaccination sentiment," Hoffman et al. (2019) identified social mistrust, safety concerns, and conspiracy ideology as key drivers, all of which flow from the *avoidance urges* coupled to *misunderstood feelings of fear*.

But to begin reclaiming the *meaningful first-person informational messages* within our emotional perceptions opens upon an entire domain of *evaluative rationality* formerly opaque to science, and provides a bulwark against the social abuses of emotion. Indeed, the binary (feel good, feel bad) nature of hedonic qualia encodes *several levels of binary logic* concerning the well-being of the self across time and social space: The most fundamental is a biologically universal "yes" or "no" evaluative logic that subserves two non-negotiable, yet potentially conflicting evolutionary *purposes* (akin to economic "utility functions"). They are subjective reflections of the imperatives for natural selection: Painful feelings demand priority *self-preservation* of the body-self in its immediate environment (Darwinian "survival" – distress signals saying "no" to self-destruction), while pleasurable feelings foster more long-term *self-development* of the mind-self (Darwinian "adaptation" – "eustress" signals [Selye 1957] saying "yes" to optimal growth, learning (including epigenetic development and neural plasticity) and culturally creative agency. A *reversal of this self-regulatory logic* gives rise to the closed-minded form of identity management exemplified in vaccine hesitancy – the exact opposite of that suggested by the logic of evolutionary utility and therefore self-destructive.

Through this lens, the resource rationality model can be enhanced by acknowledging the central role of pleasurable and painful feelings (along with their basic and complex appraisal themes) as first person informational resources, as well as dynamic capital for third-person punitive or rewarding social control, which together offer a missing dimension of *emotional reasons* that more accurately explain, predict, and might ultimately prevent behaviors like vaccine hesitancy.

Cognitively bounded rational analyses and the crucial role of theories of subjective utility

Richard L. Lewis[a] and Andrew Howes[b]

[a]Department of Psychology and Weinberg Institute for Cognitive Science, University of Michigan Ann Arbor, MI 48109 and [b]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom.
rickl@umich.edu
a.howes@bham.ac.uk
http://www-personal.umich.edu/~rickl/
http://www.cs.bham.ac.uk/~howesa/

doi:10.1017/S0140525X19001729, e14

Abstract

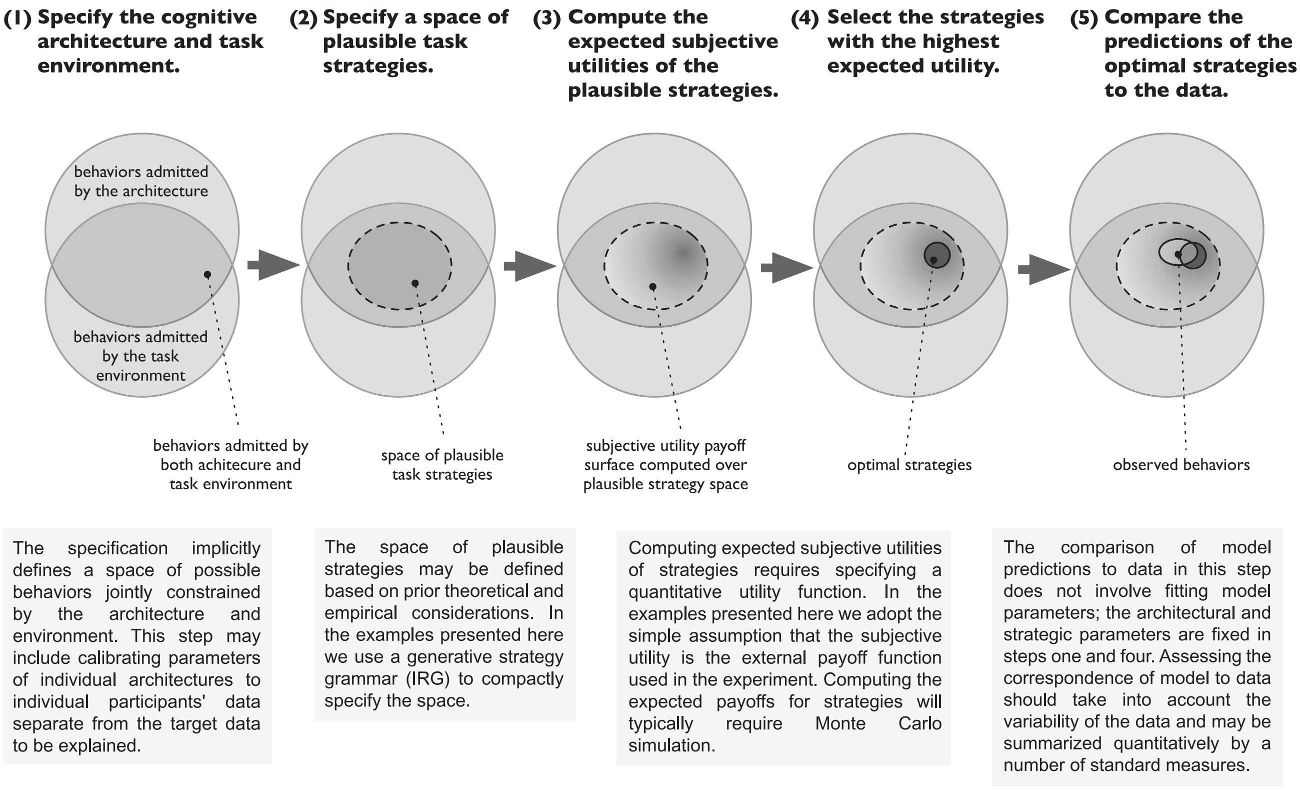
We agree that combining rational analysis with cognitive bounds, what we previously introduced as *Cognitively Bounded Rational Analysis*, is a promising and under-used methodology in psychology. We further situate the framework in the literature, and highlight the important issue of a theory of subjective utility, which is not addressed sufficiently clearly in the framework or related previous work.

The authors propose Resource Rational Analysis as a unifying modeling paradigm that combines rational analysis with cognitive constraints, arguing that it addresses the problem of underdetermination of cognitive mechanism by data, and provides a way to theorize about cognitive constraints while retaining the rigor of optimality analyses. The authors offer a multi-step method that includes a step that derives an optimal algorithm to run on the mind's computational architecture, and ground the paradigm formally in the framework of bounded optimality from artificial intelligence (AI). We endorse this paradigm and method: we made these arguments and proposed a paradigm and method in Howes et al. (2009) with just these features: *Cognitively Bounded Rational Analysis* (grounding it in bounded optimality in Lewis et al. [2014]). Each step of the authors' Resource Rational Analysis method (except for the "iterate" step) corresponds to a step in Cognitively Bounded Rational Analysis (Fig. 1), including the crucial steps that distinguish it from Anderson's seminal Rational Analysis: positing an explicit space of algorithms to run on a cognitive machine, the selection of the algorithm that maximizes some utility, and the evaluation of the optimal algorithm against data. The illustrative example used in Howes et al. (2009) has several desirable properties of Resource Rational Analysis highlighted by the authors, including a method for the automatic derivation of complex cognitive strategies (beyond optimization of quantitative parameters), and the calibration of cognitive constraints with independent data.

The target article provides a useful survey of recent relevant work across multiple domains in cognitive science. The breadth is important because it makes clear that bounded optimality analyses are useful beyond perceptual decision making and motor control, and when brought to bear on higher cognition provide new insights into the nature of human rationality. Rather than address specific applications, we focus here on an important issue that is not addressed sufficiently clearly in the authors' framework, our own previous work, or related work. The issue concerns a theory of utility.

That there is an issue can be seen in the different treatment of utility in Equations 2 and 3 in the target article. Equation 2 is a form of bounded optimality with a direct correspondence to the definitions of Russell & Subramanian (1995) and Lewis et al. (2014). The utility function is an unconstrained function of agent–environment interactions. In contrast, Equation 3 (which does not build on Equation 2) has two distinctive features: it ascribes a belief state to the agent, and it decomposes the objective function into utility and resource cost terms. We now consider some implications of this decomposition.

There was no need to include separate "cost-of-computation" or "resource-cost" terms in the original formulation of bounded optimality because these costs are captured by the implications of the machine constraints for the utility of machine–environment interactions. In particular, it is easy to specify a speed-accuracy tradeoff in a utility function that may implicitly put pressure on the machine+algorithm to make various internal trade-offs of speed, memory, accuracy, and so on. In this sense, the separate resource cost term in Equation 3 adds neither expressive power nor theoretical constraint. The incorporation of a belief state, which we take to be a probability distribution over which expectations may be computed, is a theoretical constraint. It is not clear how many of the examples reviewed in the target article actually assume an analysis that incorporates belief states; we assume it is useful in some analyses but not a commitment of the resource rationality framework.


(1) Specify the cognitive architecture and task environment.
(2) Specify a space of plausible task strategies.
(3) Compute the expected subjective utilities of the plausible strategies.
(4) Select the strategies with the highest expected utility.
(5) Compare the predictions of the optimal strategies to the data.
behaviors admitted by the architecture
behaviors admitted by the task environment
behaviors admitted by both achitecure and task environment
space of plausible task strategies
subjective utility payoff surface computed over plausible strategy space
optimal strategies
observed behaviors

The specification implicitly defines a space of possible behaviors jointly constrained by the architecture and environment. This step may include calibrating parameters of individual architectures to individual participants' data separate from the target data to be explained.

The space of plausible strategies may be defined based on prior theoretical and empirical considerations. In the examples presented here we use a generative strategy grammar (IRG) to compactly specify the space.

Computing expected subjective utilities of strategies requires specifying a quantitative utility function. In the examples presented here we adopt the simple assumption that the subjective utility is the external payoff function used in the experiment. Computing the expected payoffs for strategies will typically require Monte Carlo simulation.

The comparison of model predictions to data in this step does not involve fitting model parameters; the architectural and strategic parameters are fixed in steps one and four. Assessing the correspondence of model to data should take into account the variability of the data and may be summarized quantitatively by a number of standard measures.

Figure 1. (**Lewis & Howes**). How the five steps of cognitively bounded rational analysis focus the space of behaviors. Step 1 defines architecture and environment. Steps 2–4 narrow that space by first determining the plausible strategies and then determining the subset of best strategies. Only Step 5 involves comparison to data. From Howes et al. (2009).

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

But there is an important sense in which Equation 3 is more expressive than the bounded optimality Equation 2: It allows for the cost (and so overall utility) to be a function of the internal state of the cognitive machine. It must be if any kind of cognitive "cost" or "effort" other than time is to be calculated. The distinction is an important one; we have assumed in the past (e.g., Howes et al. 2009) that the utilities in bounded rational analyses are subjective utilities, which implies that they are functions of internal agent state. But the form of Equation 2 in the target article and our own formalization in Lewis et al. (2014) inherited from bounded optimality the property that utility is a function of states of the world/environment with which the agent interacts. We suggest that making the utility explicitly a function of internal agent state would yield a conceptually simpler and clearer definition that has the expressive power of Equation 3, without committing to a belief state formulation of agent state, or a particular kind of cost term.

These considerations put into sharp focus the need for constrained theories of subjective utility; otherwise the methodological benefits of bounded rational analysis may be diminished by the additional degrees of freedom available in specifying "resource costs." Such theories must go beyond economic models of subjective utility and include explicit accounts of cognitive effort; the author's own work on effort (Shenhav et al. 2017) begins to provide such a theory. But, more generally such theories must also encompass formal accounts of so-called "intrinsic" motivations thought to drive exploration and learning. It is in fact possible to bring a bounded-optimality analysis to bear on such theorizing: the *optimal rewards* framework (Singh et al. 2010; Sorg et al. 2010) sets up a meta-optimization problem that derives internal reward functions adapted to the bounds of learning agents so that they maximize some measure of objective fitness.

We agree that combining rational analysis with cognitive bounds is a promising and still under-used methodology for cognitive science and psychology, and the target review contributes substantially to this case. The hope is that a relatively small set of computational abstractions will emerge over time that are broadly useful as theories of cognitive mechanism. This hope looks beyond a broadly applicable method to broadly applicable theory, a hope expressed by Newell (1990) in his call for unification in psychology.

Multiple conceptions of resource rationality

Wei Ji Ma[a] and Michael Woodford[b]

[a]New York University, New York, NY 10003 and [b]Department of Economics, Columbia University, New York, NY 10027
http://www.cns.nyu.edu/malab
mw2230@columbia.edu
http://blogs.cuit.columbia.edu/mw2230/

doi:10.1017/S0140525X19001754, e15

Abstract

Resource rationality holds great promise as a unifying principle across theories in neuroscience, cognitive science, and economics. The target article clearly lays out this potential for unification. However, resource-rational models are more diverse and less easily unified than might appear from the target article. Here, we explore some of that diversity.

Resource-rational models maximize some measure of performance while simultaneously minimizing a cognitive or neural resource cost or while simultaneously satisfying a resource constraint. We observe that proposals that all start from this same high-level principle are fairly different in their implications.

To understand this diversity, we believe that it is necessary to pay careful attention to the following model dimensions.

The form of the performance term. The form of the performance term differs widely across models. Some models commit to a specific task and an associated behavioral objective, such as estimation or tracking error (Mackowiak and Wiederholt 2009; Młynarski and Hermundstad 2018; Park and Pillow 2017; Sims 2003; Sims et al. 2012; van den Berg and Ma 2018), categorization accuracy (Li et al. 2017; Młynarski and Hermundstad 2018; van den Berg and Ma 2018), or discriminability (Ganguli and Simoncelli 2014). Other models instead use mutual information between stimulus and response as a performance term (Barlow 1961; Laughlin 1981; Olshausen and Field 1996; Wei and Stocker 2015; Zaslavsky et al. 2018). The latter approach is meant to be general-purpose rather than task-specific and arguably appropriate for neural codes in early sensory areas, but suboptimal for almost any particular task. Resource-rational modelers need to make an explicit, motivated commitment to a task-specific or a general-purpose performance term, and the field needs to figure out in what situations the brain uses either type. If a model assumes a general-purpose term, its degree of suboptimality in specific tasks should be studied.

The nature of the resource cost or constraint. Similarly, there have been many different formulations of the resource cost/constraint. Some models use an information-theoretic measure of the complexity of internal representations, for example, mutual information as a cost function (Sims 2003; Sims et al. 2012), others an algorithmically motivated measure of the intensity of observation or calculation (Shaw and Shaw 1977), and yet others an explicitly neural cost (Barlow 1961; Ganguli and Simoncelli 2014; Laughlin 1981; Olshausen and Field 1996; van den Berg and Ma 2018). Another important distinction is between models that assume a cost/constraint on the number of different types of signals that can be sent, regardless of the degree to which the full repertoire is used (Laughlin 1981; Netzer 2009; Robson 2001; Steiner and Stewart 2016; Wei and Stocker 2015; Woodford 2012), and those that assume a cost/constraint on the rate at which signals are actually sent through the system (e.g., whenever the cost/constraint is on mutual information); sometimes constraints of both types are imposed (Ganguli and Simoncelli 2014). Finally, some models (such as Laughlin 1981) assume a hard constraint on the quantity of the resource that can be used, while in others (such as van den Berg and Ma 2018) the quantity of the resource is variable but there is an increasing cost of using more of it. The two formulations make different predictions about how the mechanism should be expected to change when the environment changes; the variable-resource version also allows analysis of the question of the optimal allocation of attention or precision across multiple locations or dimensions of a decision problem (Mackowiak and Wiederholt 2009; Shaw and Shaw 1977; van den Berg and Ma 2018). Beyond the fields discussed in the target article, resource-rational modelers might want to draw inspiration from the theory of optimal feedback control, in which more precise control incurs greater metabolic expenses at the organismal level (Todorov and Jordan 2002).

The time scale over which resources are allocated. Attention can be efficiently allocated in response to trial-to-trial variations in reward or priority (Bays 2014; Sims 2003; van den Berg and Ma 2018), in other words, on a timescale of seconds. By contrast, efficient neural codes are often assumed to be optimized with respect to natural statistics (Barlow 1961; Laughlin 1981), which vary on a much longer timescale. This distinction seems largely aligned with the one made under (1), with shorter timescales being associated with task specificity. Resource-rational models are often non-committal about the timescales over which the optimization occurs. Recent work on efficient codes in nonstationary environments (Młynarski and Hermundstad 2018) holds promise for bridging the divide.

Learning to be resource-rational. A question that is not often asked is how resource-rational mechanisms are learned. The target article simply defines a constrained optimum and supposes that "evolution, cognitive development, and life-long learning" have somehow solved it, without saying how. But recognizing that a particular cognitive mechanism is optimal for one's environment requires knowledge of the statistics of the environment, which in practice can never be known with certainty from any finite body of experience. The informational requirements of the learning process may impose constraints on the degree of efficiency of cognitive mechanisms that can be learned, even asymptotically, as discussed, for example, by Robson and Whitehead (2016). The question of how well-adapted a cognitive mechanism can reasonably be assumed to be is even more important if the statistics of the environment are changing (Młynarski and Hermundstad 2018).

Are finite-sampling models truly resource-rational models? In some models described in the target article, the observer simulates possible futures – technically, Markov chain Monte Carlo (MCMC) sampling from a posterior (Lieder et al. 2014; 2018; Vul et al. 2014). The high-level idea here is that samples represent computational resources, and that those are limited. More samples would correspond to a better approximation of a performance term. However, it is unclear to us if this approach falls into the framework of optimizing a linear combination of a performance term and a resource cost.

Role of reasoning. An ambiguity in references to "resource-rationality" is whether "rationality" is intended to mean the outcome of a process of conscious, logical reasoning, or simply means that something is an efficient solution to a problem, however that solution may have developed (Blume and Easley 1984; Smith 2009). Theories of efficient coding in early-stage sensory processing are rather obviously not to be interpreted as hypotheses according to which sensory processing is consciously decided upon; and it seems that in general, the authors of the target article do not have intend "rationality" in this way – the distinction that they draw between the resource-rationality hypothesis and Stigler's (Stigler 1961) model of optimal information gathering indicates this. Nonetheless, this is not clear in all of the references that they cite as examples of the resource-rationality research program. In particular, the more recent economics literature that models the imperfect information of decision makers as reflecting an optimal allocation of limited attention is often written as if the decision as to what to be aware of is made quite deliberately, just as in the work of Stigler.

We view these differences as challenges that need to be addressed but that do not invalidate the overall framework. Progress will require carefully distinguishing between the different formalisms, and finding ways to decide which ones are more applicable to particular settings.

Can resources save rationality? "Anti-Bayesian" updating in cognition and perception

Eric Mandelbaum[a], Isabel Won[c], Steven Gross[b] and Chaz Firestone[c]

[a]Baruch College, CUNY Graduate Center, Department of Philosophy, New York, NY10016; [b]Department of Philosophy, Johns Hopkins University, Baltimore, MD 21218; and [c]Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, MD 21218
emandelbaum@gc.cuny.edu
iwon1@jhu.edu
sgross11@jhu.edu
chaz@jhu.edu
http://ericmandelbaum.com
http://perception.jhu.edu
https://sites.google.com/site/grosssteven/

doi:10.1017/S0140525X19001717, e16

Abstract

Resource rationality may explain suboptimal patterns of reasoning; but what of "anti-Bayesian" effects where the mind updates in a direction *opposite* the one it should? We present two phenomena – belief polarization and the size-weight illusion – that are not obviously explained by performance- or resource-based constraints, nor by the authors' brief discussion of reference repulsion. Can resource rationality accommodate them?

Resource rationality takes seemingly irrational behaviors and reframes them as rational or optimal given other constraints on agents. For example, anchoring-and-adjustment and overestimating extreme events turn out be "rational" after all, by reflecting the rational *allocation* of cognitive resources. Thus, even for such classically irrational phenomena, "the resulting train of thought eventually converges to the Bayes-optimal inference" (p. 38).

In such cases, reasoners *fall short* of perfectly rational updating, and it is illuminating that resource- and performance-based constraints can accommodate such suboptimal reasoning. But what about cases where we behave not merely suboptimally, but rather *against* the norms of Bayesian inference? Here, we explore cases where the mind is moved by prior knowledge in precisely the *reverse* direction of what a rational analysis would recommend. These cases are not merely suboptimal, but rather "*anti*-Bayesian," for actively defying Bayesian norms of inference. We consider two such phenomena: belief polarization and sensory integration (Fig. 1). Can resource rationality handle them?

First, belief polarization: Receiving evidence contrary to your beliefs should soften those beliefs, even if ever-so-slightly. But, this isn't what actually happens when the beliefs in question are central to one's identity – in belief polarization, contrary or disconfirming evidence causes more *extreme* beliefs, not more moderate ones. A classic example was vividly documented by Festinger et al. (1956): Cult members who predict the world will end on some date – but who then see that date come and go with no

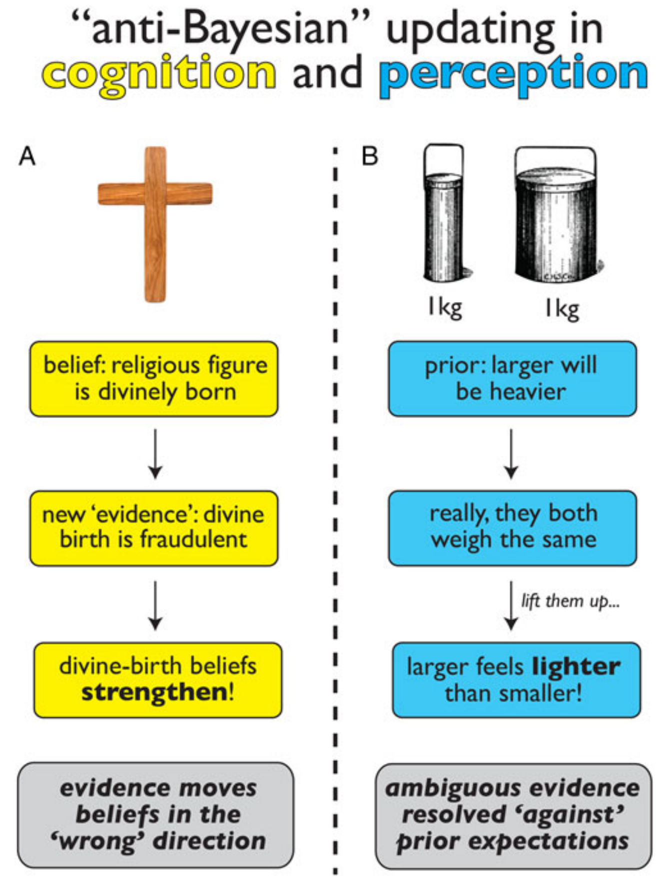

"anti-Bayesian" updating in
cognition and perception
A
belief: religious figure is divinely born
new 'evidence': divine birth is fraudulent
divine-birth beliefs strengthen!
evidence moves beliefs in the 'wrong' direction
B
1kg
1kg
prior: larger will be heavier
really, they both weigh the same
lift them up...
larger feels lighter than smaller!
ambiguous evidence resolved 'against' prior expectations

Figure 1. (**Mandelbaum et al.**) Examples of "anti-Bayesian" updating in the mind. (A) Under conditions of cognitive dissonance, acquiring – and affirming – evidence *against* one's beliefs can cause those beliefs to strengthen (Batson 1975), whereas Bayesian norms of inference recommend softening those beliefs. (B) In the size-weight illusion, one is shown two objects of different sizes but equal weights; when one lifts them up, the *smaller* one feels illusorily heavier than the larger one (Buckingham 2014; Charpentier 1891; Won et al. 2019). In other words, ambiguous sensory data about which of two objects is heavier is resolved "against" one's prior expectations, rather than in favor of one's priors as recommended by Bayesian norms of inference. Can resource rationality accommodate such paradigmatically "irrational" phenomena?

cataclysm – end up *strengthening* their beliefs in the cult's tenets, not softening them. In other words, credible evidence *against* their worldview only makes them hold that worldview more strongly – directly defying Bayesian inference norms.

The same phenomenon can be found under laboratory conditions. For example, one study exposed people who believe that Jesus is the Son of God to a (fake) news article reporting that archeologists had unearthed carbon-dated letters from the New-Testament authors; the letters said the Bible was fraudulent and that its authors knew Jesus was not divinely born (Batson 1975). Subjects who did not believe the article's content left their beliefs about Jesus unchanged; but, fascinatingly, subjects who did believe the article's content ended up *strengthening* their belief that Jesus was the Son of God. In other words, *affirming* new evidence *against* Jesus's divine birth (~*P*) caused stronger beliefs in Jesus's divine birth (*P*). Similar "backwards" updating is also observed for beliefs about nuclear safety (Plous 1991), health (Liberman & Chaiken 1992), and affirmative action and gun control (Taber & Lodge 2006; see also Mandelbaum 2019).

Why does this happen? In fact, belief polarization is not so mysterious: It has been known for decades, and it is even a predictable consequence of dissonance theory – "the psychological immune system" (Gilbert et al. 1998) – applied to one's values. What *is* mysterious is why this should occur *in a Bayesian mind* – even one constrained by "resources." Belief polarization is irrational not because people are *insufficiently moved* by evidence, but rather because people are moved in the direction *opposite* the one they should be. And, importantly, these patterns cannot be explained by biased attitudes toward the evidence's source. For example, Bayesian models of milder forms of belief polarization (e.g., Jern et al. 2014) suggest that subjects infer that contrary evidence must have come from unreliable sources (e.g., biased testimony); but this seems inapplicable to the above cases, where the sources are either nature itself (e.g., the world failing to end), or evidence the subject has actively accepted (e.g., news articles they endorsed).

Indeed, "anti-Bayesian" updating is widespread, occurring even in basic perceptual processes. When we have prior expectations about new and uncertain sensory data, rational norms of inference say we should interpret such data with respect for those priors; "people should leverage their prior knowledge about the statistics of the world to resolve perceptual uncertainty" (p. 40). But, sensory integration frequently occurs the opposite way. Consider the size-weight illusion, wherein subjects see two equally weighted objects – one large and one small – and then lift them both to feel their weight. Which feels heavier? We "should" resolve the ambiguous haptic evidence about which object is heavier *in favor* of our priors; but instead, the classic and much-replicated finding is that we experience the smaller object as *heavier* than the equally-weighted larger object (Buckingham 2014; Charpentier 1891). This too is "irrational" – not for *falling short* of Bayesian norms of inference, but for proceeding opposite to them, because we resolve the ambiguous sensory evidence – two equally weighted objects – *against* the larger-is-heavier prior, not in favor of it (Brayanov & Smith 2010; Buckingham & Goodale 2013). Indeed, this backwards pattern of updating is so strong that it can produce outcomes that are not merely odd or improbable, but even "impossible" (Won et al. 2019): If subjects are shown three boxes in a stack – Boxes A, B, and C – such that Box A is heavy (250 g) but Boxes B and C are light (30 g), then subjects who lift Box A alone and then Boxes A+B+C together report that Box A feels heavier than Boxes A+B+C – an "impossible" experience of weight (because a group could never weigh less than a *member of that group*).

How can a "rational" account – even a resource-rational one – explain this? Lieder and Griffiths accommodate other sensory "repulsion" effects (Wei & Stocker 2015; 2017), but that modeling work seems inapplicable to the size-weight illusion. And whereas the original size-weight illusion could perhaps have a tortuous Bayesian explanation (Peters et al. 2016), Won et al.'s modification seemingly cannot: First, it's unclear if previous models of simultaneous lifting apply to Won et al.'s temporally-extended case; but second, there is just no logical chain of reasoning that should end with A alone being heavier than A+B+C together.

More generally: What are the principles that lead to perverse "anti-Bayesian" updating? Perhaps resource rationality wasn't intended to cover all cases (in which case it is not an "Imperial Bayesian" theory; Mandelbaum 2019). But, the problem isn't merely that there are counterexamples to resource rationality, but rather that these are predictable, law-like counterexamples that do not reflect performance constraints between interacting mental processes. Indeed, when it comes to these more entrenched patterns, even "resources" may not save rationality.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Towards a quantum-like cognitive architecture for decision-making

Catarina Moreira[1], Lauren Fell[1], Shahram Dehdashti[1], Peter Bruza[1] and Andreas Wichert[2]

[1]School of Information Systems, Science and Engineering Faculty, Queensland University of Technology, Brisbane City QLD 4000, Australia and [2]Instituto Superior Técnico, University of Lisbon/INESC-ID, 2744-016 Porto Salvo, Portugal.
catarina.pintomoreira@qut.edu.au
https://staff.qut.edu.au/staff/catarina.pintomoreira
l3.fell@qut.edu.au
https://staff.qut.edu.au/staff/l3.fell
shahram.dehdashti@qut.edu.au
https://qutvirtual4.qut.edu.au/web/qut/person-details?id=01783299&roleCode=EMPp.bruza@qut.edu.au
https://staff.qut.edu.au/staff/p.bruza
andreas.wichert@tecnico.ulisboa.pt
http://web.tecnico.ulisboa.pt/~andreas.wichert/

doi:10.1017/S0140525X19001687, e17

Abstract

We propose an alternative and unifying framework for decision-making that, by using quantum mechanics, provides more generalised cognitive and decision models with the ability to represent more information compared to classical models. This framework can accommodate and predict several cognitive biases reported in Lieder & Griffiths without heavy reliance on heuristics or on assumptions of the computational resources of the mind.

Lieder and Griffiths (L&G) propose a normative bounded resource-rational heuristic function to relax the optimality criteria of the expected utility theory and justify the choices that lead to less optimal decisions. Expected utility theory and classical probabilities tell us what people should do if employing traditionally rational thought, but do not tell us what people do in reality (Machina 2009). Under this principle, L&G propose an architecture for cognition that can serve as an intermediary layer between neuroscience and computation. Whilst instances where large expenditures of cognitive resources occur are theoretically alluded to, the model primarily assumes a preference for fast, heuristic-based processing. We argue that one can go beyond heuristics and the relaxation of normative theories like the expected utility theory, in order to obtain a unifying framework for decision-making.

The proposed alternative and unifying approach is based on a quantum-like cognitive framework for decision-making, which not only has the ability to accommodate several paradoxical human decision scenarios along with more traditionally rational thought processes, but can also integrate several domains of the literature (such as artificial intelligence, physics, psychology and neuroscience) into a single and flexible mathematical framework.

Figure 1 presents the proposed unifying quantum-like framework for decision-making grounded in the mathematical principles of quantum mechanics without the specification of heuristics to accommodate the cognitive biases addressed by L&G.

In quantum cognitive models, events are represented as multi-dimensional vectors according to a basis in complex Hilbert spaces, which reflects the potentials of all events. In quantum mechanics, this property refers to the *superposition* principle. For instance, when making a judgement whether or not to buy a car, a person is *at the same time* in an indefinite state corresponding to *buy* and, in the state, to *not buy* (Figure 2). Each person reasons according to their own basis. Different personal beliefs are obtained by rotating their basis, leading to different

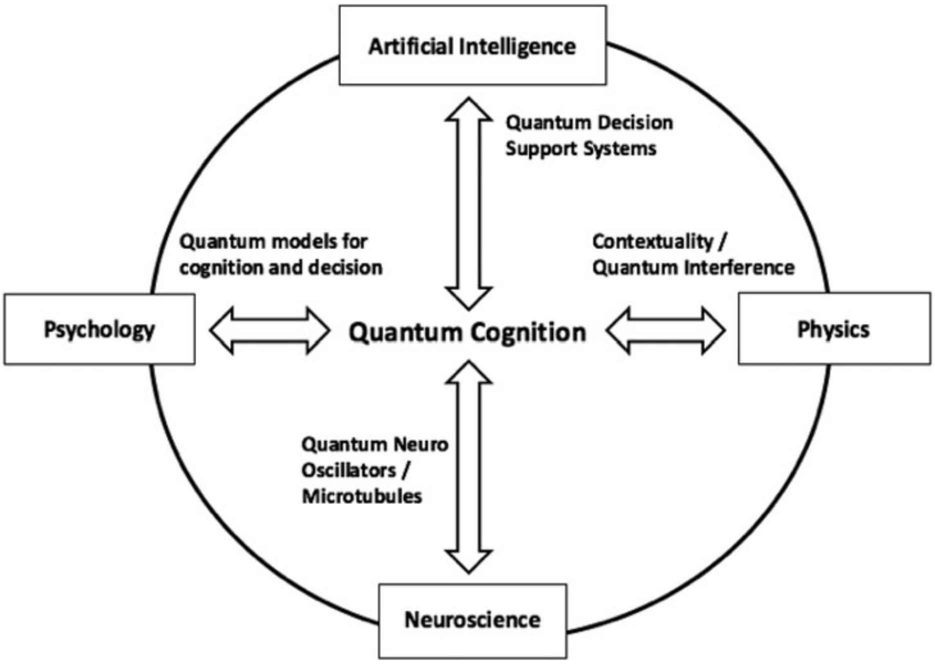

Artificial Intelligence
Quantum Decision Support Systems
Quantum models for cognition and decision
Contextuality / Quantum Interference
Psychology
Quantum Cognition
Physics
Quantum Neuro Oscillators / Microtubules
Neuroscience

Figure 1. (Moreira et al.) A quantum-like cognitive architecture

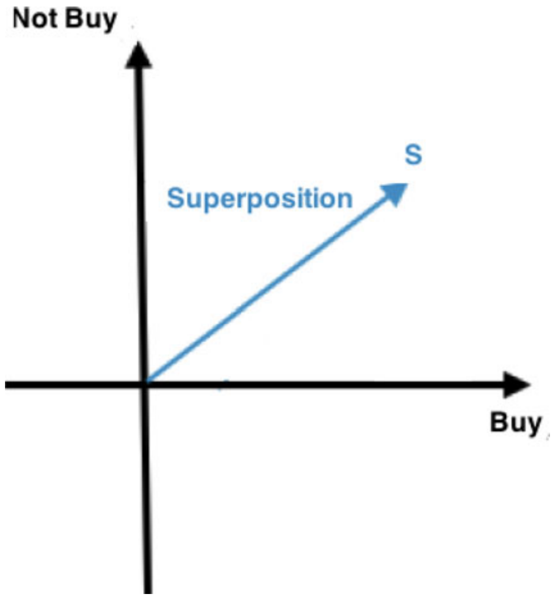


Figure 2. (Moreira et al.) Hilbert space representation of a basis state

representations of the decision scenario and different beliefs according to each decision-maker (Figure 3). Ultimately, a person can be in a superposition of thoughts in a *wave-like structure*. This can create *interference effects* leading to outcomes that cannot be predicted by classical theories. Interference is one of the core concepts in quantum cognition.

This wave-like paradigm enables the representation of conflicting, ambiguous, and uncertain thoughts (Busemeyer and Bruza 2012) and also undecidability (Cubitt et al. 2015). The superposition vector representation obeys neither the distributive axiom of Boolean logic nor the law of total probability. As a consequence, it enables the construction of more general models that cannot be captured in traditional classical models. The *accessibility of information in the quantum cognitive framework is much higher* than in a classical system, making it possible to model the different minds of bounded cognitive resources. This additional information can also accommodate several paradoxical decision scenarios, generate novel non-reductive models of how humans process concepts and generate new understandings of human cognition (Frauchiger and Renner 2018; Vourdas 2019). These distinctive features of quantum theory provide several advantages and more accurate and elegant explanations for empirical data in situations where classical probability theory alone leads to puzzling and counterintuitive predictions (cognitive bias, order effects, conjunction/disjunction errors, and so on). Although, classical probability traditionally assumes independence of events, quantum theory provides probabilistic inferences which are strongly context dependent: the same predicate may appear plausible or not, depending upon the decision-maker's point of view (Pothos and Busemeyer 2013).

L&G make a strong assumption that the mind is a computational architecture, which uses certain classes of algorithms that make the trade-off between the computational cost of using the mind's resources (and getting the necessary information) and the utility of finding the correct solution of a cognitive problem (specified at a computational level). The proposed quantum cognitive framework imposes no such assumptions about the human mind and rests on two important aspects of quantum mechanics: *contextuality* and *interference*.

Contextuality entails the "impossibility of assigning a single random variable to represent the outcomes of the same measurement procedure in different measurement conditions" (de Barros and Oas 2016, p. 153). As a consequence, it is not possible to define a single joint probability distribution from the empirical data collected from different measurement conditions such that the empirical data can be recovered by marginalising the joint distribution. Recent empirical evidence suggests that contextuality manifests in cognitive information processing (Basieva et al. 2019; Cervantes & Dzhafarov 2018). Should contextuality be present, then it would call into question the assumption of the distribution $P(result|s0,h,E)$ (Equation 4). The intuition here is that the cognitive agent cannot form this distribution because the functional identity of the random variable of which "result" is an outcome is *not* unique across the environments E (viewed as measurement conditions in regard to the quote above) (Dzhafarov & Kujala 2014; 2016). Although it is theoretically speculative to associate quantum-like contextuality with (Equation 4), we do so to draw attention the fact that contextuality has little known and undiagnosed consequences for the development of probabilistic models in cognitive science.

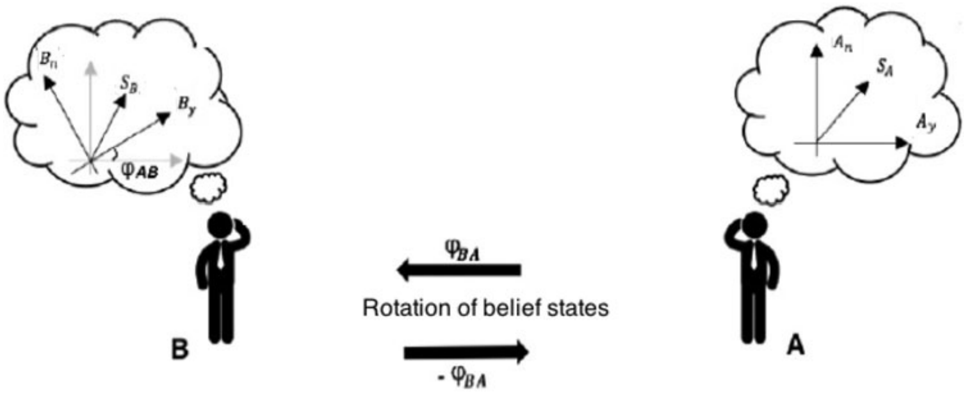


Figure 3. (Moreira et al.) Each person reasons according to an *N*-dimensional vector space by rotating their basis state towards their personal beliefs.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Interference has a major impact on cognitive models, because in decision scenarios under uncertainty, one can manipulate these quantum interference effects to disturb classical probabilities and, consequently accommodate most of the cognitive limitations reported in L&G and provide alternative inferences under uncertainty that are not captured by classical probability (Pothos and Busemeyer 2009). Recent studies suggest that quantum interference can be used to model real-world financial scenarios with high levels of uncertainty, showing a promising approach for decision support systems and artificial intelligence (Moreira et al. 2018). In Neuroscience, interference effects in the brain can occur if neuronal membrane potentials have wave-like properties (de Barros & Suppes 2009).

To conclude, we commented on the architecture proposed by L&G, which assumes a preference for fast, heuristic-based processing and strong computational assumptions about the human mind. We proposed a unifying framework for decision-making based on quantum mechanics that provides more generalised decision models capable of representing more information than classical models. It can accommodate several paradoxical findings and cognitive biases and lead to alternative and powerful inferences focused on the perspective-dependency of the decision-maker.

Acknowledgments. Dr. Andreas Wichert was supported by funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/2019. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Opportunities and challenges integrating resource-rational analysis with developmental perspectives

Kimele Persaud, Ilona Bass, Joseph Colantonio, Carla Macias and Elizabeth Bonawitz

Department of Psychology, Rutgers University–Newark, Newark, NJ 07103.
kimele.persaud@rutgers.edu
ilona.m.bass@gmail.com
jac621@scarletmail.rutgers.edu
cm1172@scarletmail.rutgers.edu
lbaraff@gmail.com

doi:10.1017/S0140525X19001560, e18

Abstract

Lieder and Griffiths present the computational framework "resource-rational analysis" to address the reverse-engineering problem in cognition. Here we discuss how developmental psychology affords a unique and critical opportunity to employ this framework, but which is overlooked in this piece. We describe how developmental change provides an avenue for ongoing work as well as inspiration for expansion of the resource-rational approach.

The power of any modeling and analysis approach comes with the degree to which it can speak to, and be informed by, variability. Cognitive development provides a rich source of variability in representation, constraint, and mechanism. This affords a unique opportunity to explore heuristic differences predicted by resource-rational analysis. Below we focus on four areas of development, detailing how each integrates with the resource-rational framework and provides a critical test of the approach.

Lieder and Griffiths' resource-rational analysis combines rational principles with cognitive constraints. Under this framework, the cost of various heuristics should be sensitive to the structure of the cognitive representations on which they operate. For example, carrying out a specific heuristic (e.g., for categorical inference) could be more costly under certain cognitive representations (e.g., non-overlapping clusters) than others (e.g., taxonomic ones). This representational variability provides a critical test of resource-rational models. Although the structure of representations is domain-dependent, cognitive representations within a domain can vary. This variability arises from development (Chi & Ceci 1987; Kemp & Tenenbaum 2008). Because heuristic cost depends on the representational structure, changing heuristics and the variability in these representations throughout development can inform the robustness and flexibility of the resource-rational approach.

Cost and availability of heuristics (and therefore their utility) in Lieder and Griffiths' framework are also influenced by cognitive constraints. Constraints of working memory capacity, executive function, and inhibition change developmentally (Davidson et al. 2006), as does the trade-off between cognitive flexibility (e.g., rule switching), recall accuracy, and processing speed (Crone et al. 2006). This suggests another important avenue for applying the resource-rational framework in development to investigate resulting changes in cognitive heuristics. For example, when older children are presented with an increase in cognitive load (e.g., increased inhibition demands), they display an increase in reaction time and higher recall accuracy, whereas younger children maintain their reaction time at the expense of recall accuracy (Davidson et al. 2006). Therefore, the variability seen in cognitive control and flexibility across development, and the implications it has on the duration and execution of cognitive computations and decisions, makes this a promising domain of research to explore the resource-rational framework.

Emotional and motivational states, factors of "internal" environment, are also critical to the cost function, as well as the availability of heuristics considered. Such affective states can influence information processing strategies adopted by individuals, assuming that cognition adapts itself to contextual requirements (known as the "feelings-as-information" perspective; Schwarz 2002; Schwarz and Clore 2007). For example, past research suggests that most negative states (e.g., sadness or fear) typically signal problems that foster systematic, bottom-up processing with attention to detail, adaptive to goal-directed behavior (Wegner & Vallacher 1986). In contrast, many positive states (e.g., happiness) are associated with reliance on heuristics and top-down use of pre-existing knowledge structures (Bless et al. 1996; Griskevicius et al. 2010). Given that children are perhaps the most variable emoters (Lewis 2008), this provides another unique opportunity of high variance to employ this analysis, especially as research begins to develop new theory integrating development with the domains of emotion and cognition (Calkins & Bell 2010).

Finally, and perhaps most critically, variability in early environmental experiences may be particularly informative because it will shape how cost functions are learned and govern which heuristics are more readily employed. At the broad level of development, for instance, theories suggest that the relative security of a protected childhood changes costs associated with "riskier" cognitive exploration in adulthood (Gopnik et al. 2017). Individual differences may also critically influence acquired cost functions – for example, recent work suggests that the kinds of questions parents tend to ask their children (Yu et al. 2019), as well as the quality of explanations parents provide in response to their children's questions (Kurkul & Corriveau 2018), systematically vary with several key factors of home life. A child whose parents are less likely to ask questions or provide causal explanations may thus acquire a very different-looking cost function for (e.g.,) the heuristic of reaching out to others for information than a child whose parents are more likely to engage in these kinds of behaviors. Indeed, this notion is consistent with recent computational work which suggests that learners may bring expectations about the teaching style of their informant to bear in future learning (Bass et al. 2018).

Although development provides special opportunities to employ resource-rational analysis by leveraging variability in the population, challenges remain. First, the *goals* of a developing system may radically vary from those in adulthood. For example, the goals of an adult semantic memory system might be defined by compression and storage for optimal later accessibility (e.g., Anderson 1989); however, hypothetically, a developing memory system's goal might be to expand and re-encode for representational restructuring. Because there is significantly less work that has focused on defining goals of the developing mind, resource-rational models will be underconstrained.

Second, variability within a developing child presents a challenge as algorithmic utilities are learned. According to the rational-resource analysis, the max ordered value of a heuristic depends on utilities that will be derived from representation, cognitive constraints, experiences, and rule-discovery. But these are constantly shifting in development, so how might a learner develop a preference for a particular heuristic? Consider a learner whose working memory limitations lead to favoring a "local search" heuristic. Although the learner's working memory capacity may grow over time, once a particular heuristic has been learned and habitually adopted, it is not clear when or why the system would re-evaluate and discover a more optimal "global" search heuristic employing newly developed resources. Such considerations suggest that a broader, dynamic framework of resource-rational analysis will need to be developed.

Overall, we think the resource-rational approach presented by Leider and Griffiths will be an important computational toolkit for cognitive psychology. Although there are challenges, we suggest that the variability found in cognitive development in particular will be a critical playground for modelers employing this technique.

Resource-rational analysis versus resource-rational humans

Dobromir Rahnev

School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332.
rahnev@psych.gatech.edu
www.rahnevlab.gatech.edu

doi:10.1017/S0140525X19001699, e19

Abstract

Lieder and Griffiths advocate for resource-rational analysis as a methodological device employed by the experimenter. However, at times this methodological device appears to morph into the substantive claim that humans are actually resource-rational. Such morphing is problematic; the methodological approach used by the experimenter and claims about the nature of human behavior ought to be kept completely separate.

A healthy adult asked to run 60 m will likely sprint; a healthy adult asked to run 1,000 m will likely jog. In fact, there is hardly anyone on Earth who would even attempt to sprint for 1,000 m. This simple observation demonstrates what we intuitively already know: that human behavior is typically adapted to our own limitations. Therefore, a deep understanding of behavior necessitates that various sources of limitations are rigorously identified and precisely quantified. I applaud Lieder and Griffiths (L&G) for advocating for this practice.

L&G propose "resource-rational analysis," which is a methodological device that an experimenter uses to discover something about human behavior. However, the target article appears to sometimes conflate this methodological device with the substantive claim that humans are actually resource-rational. To be fair, L&G stop short of claiming that people are actually resource-rational. They even offer that "we should not expect people's heuristics to be perfectly resource-rational" (sect. 3, para. 6). Nevertheless, other parts of the target article give a sense that L&G really do think that people are (mostly) resource-rational. For example, they consider seriously the "assumption that the brain is approximately bounded-optimal" (sect. 5.3.2., para. 2), claim that resource-rational analysis "has already shed new light on the debate about human rationality" (abstract), and even state that "people's decision-mechanisms appear to be surprisingly resource-rational" (sect. 6, para. 3). These statements leave the realm of methodological devices and venture into the land of substantive claims about human rationality.

The problem is that, as currently constructed, resource-rational analysis does not and could not provide evidence for the rationality of human behavior. There are at least three reasons for this.

First, resource-rational analysis in overly flexible as a tool for establishing the nature of human behavior. As Box 2 demonstrates, a researcher who follows the methodology prescribed by L&G should test a number of different constraints and computational architectures until some combination of them provides a good fit to the data. To L&G's credit, they do advice that the experimenter stops trying out new combinations after "reasonable attempts have been made to model the constraints" (Box 2). Nevertheless, for most experimental tasks, it is not too difficult to find a set of assumptions that makes behavior to appear close to rational. This does not, however, imply that the underlying behavior is rational because the experimenter may have unwittingly postulated computational architectures or resource limitations that do not exist, or, more likely, exist but are mischaracterized. For example, a tendency to underuse explicitly stated probabilities (Rahnev and Denison 2018a) can be cast as optimal decision making by an organism that misrepresents probabilities (Zhang and Maloney 2012). However, this explanation could be given regardless of whether the organism actually adopts optimal decision making based on skewed representations of probability or adopts a suboptimal decision strategy on internal representations of probability that are less skewed. Therefore, substantive claims about human rationality require models that are prespecified and have no free parameters (e.g., the misrepresentation of probabilities should be predetermined for each subject). Very few papers, however, fit such zero-parameter models to the data.

Second, the types of tasks that we study in the laboratory tend to be the most constrained and simple tasks that an organism could ever face. Yet, even for such tasks, suboptimality is the norm (Rahnev and Denison 2018a). Regardless of how close to rationality humans get in such tasks, it does not follow that behavior would be similarly rational in the infinitely more complex real world. As L&G admit themselves, it is "challenging to [apply resource-rational analysis] to decision-making in the real world where the sets of options and possible outcomes are much larger and often unknown" (sect. 6, para. 7).

Third, the computations required to establish the truly rational strategy are intractable and will always remain so. Indeed, as demonstrated by Equation 4 in the target article, specifying what is actually rational requires quantitatively describing all environments that one has ever experienced (including environments that have been experienced by one's ancestors and have influenced brain development over evolutionary scales), which is clearly infeasible in practice. Therefore, in the strictest sense of Equation 4, we will never be able to test whether any behavior is truly rational or not.

If there is little hope that we could ever establish whether human behavior is really rational, does that mean that resource-rational analysis is also futile? Not at all. As the example of running a shorter versus longer distance demonstrates, we are profoundly constrained by our limitations, and our behavior is often roughly adapted to these limitations. Therefore, resource-rational analysis offers at least two large benefits (in addition to what was highlighted by L&G). First, resource-rational analysis can be used to *approximate* human behavior under the assumption that evolution has adapted our behavior to the particular task used by the experimenter. Clearly, for a non-resource-rational human, the approximation may be crude and sometimes very imprecise, but at the very least could be used as a starting point. Second, behavior that is systematically deviating from resource-rationality may indicate the existence of a new, previously undiscovered limitation or cognitive architecture. As highlighted above, postulating limitations just for the sake of fitting data is a dangerous undertaking, and thus any proposal for a new limitation should be tested with independent data and, ideally, under new conditions.

Regardless of one's preferred view of human nature and the best methods to reveal that human nature, it is critical that substantive claims about behavior and methodological approaches about studying said behavior are kept separate from each other (Rahnev and Denison 2018b). The person who jogs for 1,000 m is unlikely to do so at the optimal pace. That is, she is unlikely to be fully resource-rational. However, we will certainly understand her behavior better if we put in the effort to quantify the exact rate at which her muscles tire. Resource-rational analysis can be useful even if we are trying to characterize non-resource-rational humans.

Resource-rationality and dynamic coupling of brains and social environments

Don Ross[a,b,c]

[a]School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork, Cork T12 AW89, Ireland; [b]School of Economics, University of Cape Town, Rondebosch 7701, South Africa and [c]Center for Economic Analysis of Risk, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303.
don.ross931@gmail.com
http://uct.academia.edu/DonRoss

doi:10.1017/S0140525X1900150X, e20

Abstract

Leider and Griffiths clarify the basis for unification between mechanism-driven and solution-driven disciplines and methodologies in cognitive science. But, two outstanding issues arise for their model of resource-rationality: human brains co-process information with their environments, rather than merely adapt to them; and this is expressed in methodological differences between disciplines that complicate Leider and Griffiths' proposed structural unification.

Leider and Griffiths' (L&G) project, to offer an explicit framework for relativizing assessments of rationality simultaneously to cognitive processing constraints and environmental affordances, represents important progress. It significantly clarifies the basis for unification between mechanism-driven and solution-driven disciplines and methodologies, as they say. But, as the framework is extended and refined, two outstanding issues merit consideration: (1) human brains do not merely adapt to their environments, but co-process information with their environments, particularly with its social aspects; and (2) L&G's idealization of disciplines as standing in a hierarchy of abstraction from mechanism details is a somewhat misleading simplification of methodological reality.

L&G's core Equation 4 takes the environment (E) as a fixed constraint on optimal heuristic selection. This is reasonable in light of the long time-scale for learning that their discussion indicates that they have in mind, reflected in their comment that evolution and cognitive development "solve the constrained optimization problem defined in Equation 3" (sect. 3, para. 5). The framework obviously allows for environmental variation, across time or space, to be modeled and analyzed using comparative statics. Furthermore, their inclusion of the information term I on the left-hand side of Equation 4 recognizes that learning encoded in the genome is refined by learning in the phenome. However, the model seems to presuppose that cognitive processing is all done in the brain, because there is no interaction term involving all of h, E, and B (heuristics, environment, and brain).

This may be a reasonable idealization where most cognitive systems are concerned. But, it might be seriously misleading in the case of humans equipped with writing, art, and mathematics, who have populated their environments with technologies that actively process information in conjunction with inboard cognition. Obvious examples include external computing devices, but these are not the main source of potential deep complication for L&G's model. Though, the relationship between a person and a machine she uses may be dynamically interactive, in non-exotic cases the extent of such dynamical coupling is both limited and specifiable; and, as noted above, this is all that is required for analyzing variation by means of comparative statics. The more serious challenge arises from the abstract technology of social institutions. Ecologically, humans are arguably most strongly distinguished from other highly intelligent animals by their use of shared information-processing routines that are encoded in rules, norms, and institutionalized procedures, which individuals exploit by mixtures of faithful and noisy compliance and deliberate modification. It is not immediately evident how L&G's Equation 4 should be modified to parcel elements of such social heuristic (or optimizing) processes into inboard and outboard elements. Put in terms of a simple example: if an entrepreneur generally follows her venture capitalist's boilerplate advice, but distorts it through a mix of subjective probability weighting on risks and explicit private knowledge, and this in turn influences all participants' models of their market, which elements are to be included in *B* and which in *E*?

L&G quote Simon's salutary observation that "the environment may lie, in part, within the skin of the biological organism" (sect. 3, para. 7). Their model reflects this insight. But, then a complementary point, as emphasized by theorists such as Clark (1997) and Sterelny (2003; 2012), is that some cognitive processing occurs outside of the skin. I stress again that I am not referring here to information processing that is largely exogenous to psychological mechanisms, as in the computation of graphical representations by a statistics package or of prices by a market. To call such processing "cognitive" would involve capture by a metaphor. The point, rather, is that as Clark (2003) emphasizes, people expand their intelligence by coupling their brains with the representational and active computational tools that they collectively operate; and the boundary between individual and social resource-rationality is, in the context of conserved engineering achievements, blurred to the extent of collapse. Reliance on outboard processing for much of the very sophisticated information processing characteristic of humans is plausibly an essential requirement arising from metabolic constraints.

This concern is related to an assumption L&G make explicit later in their paper about the relationships among disciplines. They describe economics (along with AI) as simplifying and idealizing models of the mind. Behavioral economists who seek strong unification between their discipline and psychology are likely to be comfortable with this, but it obscures important methodological differences. Psychometrics is, to a first approximation, the statistical theory of measures of construct validity because psychologists aim to infer 'hidden' mechanisms from observations of behavior, and thus need to exclude 'confounding' elements of *E* when experimentally focused on *B*. By contrast, experimental economists tend to deliberately undermine the importance of the *E*/*B* distinction by adding new treatments where psychologists would seek to block out a "confound." This explains why the econometrics of the lab is essentially the statistical theory of structural model identification and estimation. Although, economists are of course students of information-processing, it does not seem apt to depict them as studying abstract, idealized minds. This might look like merely philosophical quibbling. But, in fact it anchors the concern about the clarity of the *E*/*B* distinction in terms of practical modeling: as an economist I would need detailed reassurance that if standard, working structural models are to be constrained by Equation 4, this would not require solving identification problems that economists have worked hard to bypass.

I close with an example. Cumulative prospect theory, which is certainly an idealized model of mind just as L&G say it is, is an awkward tool for the economist's lab because it locks in parameters that are extremely difficult to identify (Harrison and Swarthout 2016), and which lack principled theoretical generalization. And almost all of the relevant empirical estimation work can instead be done using a rank-dependent utility specification that allows for subjective decision weights that fail to track objective probabilities. Do the subjective weightings in question come from *B* or from (the social) *E* (see Harrison and Ross 2017)? Does the economist really need to care about the answer?

Opportunities for emotion and mental health research in the resource-rationality framework

Evan M. Russek[a,b], Rani Moran[a,b], Daniel McNamee[a,b], Andrea Reiter[a,b], Yunzhe Liu[a,b], Raymond J. Dolan[a,b] and Quentin J.M. Huys[a,c,d]

[a]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, UK; [b]The Wellcome Trust Centre for Neuroimaging, London, WC1N 3AR, UK; [c]Division of Psychiatry, University College London, London, W1T 7NF, UK and [d]Camden and Islington NHS Foundation Trust, London, NW1 0PE United Kingdom.
e.russek@ucl.ac.uk
r.moran@ucl.ac.uk
d.mcnamee@ucl.ac.uk
yunzhe.liu.16@ucl.ac.uk
a.reiter@ucl.ac.uk
r.dolan@ucl.ac.uk
q.huys@ucl.ac.uk

doi:10.1017/S0140525X19001663, e21

Abstract

We discuss opportunities in applying the resource-rationality framework toward answering questions in emotion and mental health research. These opportunities rely on characterization of individual differences in cognitive strategies; an endeavor that may be at odds with the normative approach outlined in the target article. We consider ways individual differences might enter the framework and the translational opportunities offered by each.

The resource-rationality framework presented by Lieder and Griffiths has the potential to open up new computational approaches to emotion, including in the setting of mental illness. However, pitfalls can arise from a strong emphasis on normative modeling when this is at the expense of approaches that allow for measurement of individual differences. We consider the latter important in many translation efforts.

Minds face extremely complex decision-making tasks that far outstrip available computational resources. An individual must decide how to optimally apportion cognitive resources, such as attention, what memories to prioritize, what actions to evaluate and in what order, what future situations to simulate, whether to expend or save energy resources, and so on. The computational burdens imposed by resource-rationality, to choose optimal internal actions, are huge and often mandate the use of fast, automatic, and approximate solutions.

Emotions and moods have long been characterized as states involving coordinated biases in all these domains (e.g. Scherer 2009), suggesting that they could function as psychological and physiological mechanisms by which humans implement approximate resource-rational computations (Huys & Renz 2017). For example, different emotions prioritize distinct action sets for mental evaluation (Frijda et al. 1989); an effect which can be interpreted as implementing a solution to the meta-cognitive problem of the order in which a large space of actions should be evaluated. Moods involve longer-lasting biases toward particular emotions, and hence result in more persistent sets of cognitive choices to such internal demands. For example, low mood states are characterized by preferentially choosing to attend to negative stimuli, a prioritization of negative memories, a preferential evaluation of avoidance actions and not expending energy.

If it is true that moods are coherent, systematically covarying sets of internal cognitive choices that approximately address resource allocation problems, the question arises why a particular set of cognitive choices that define a mood state tend to co-occur. Furthermore, how these internal policies change and adapt becomes fundamental to understanding them. The framework of resource-rationality seems ideally placed in answering these questions. By defining what state features lead a resource rational agent to make the cognitive selections defined by a given mood, we can understand why particular cognitive actions become associated with one another in distinct meta-action sets, as well as what pieces of environmental information trigger mood shifts, and hence why the range of internal actions becomes partitioned into a particular set of emotions.

Cognitive action selections are likely aberrant in many mental illnesses. Most obviously, if emotions are indeed cognitive action policies, this suggests that mood disorders can arise from maladaptive solutions to internal resource allocation problems. Indeed, patients with depression are known to suffer from characteristic biases in their allocation of attention, memory recall, action, and energy expenditure (Elliott et al. 2011; Mathews & MacLeod 2005; Whitton et al. 2015). Similarly, a tendency to catastrophize might relate to resource-rational arguments for biasing mental simulation toward extreme events (Lieder et al. 2018). However, this counter-intuitively links normative functions to maladaptive psychopathology, and begs the question about individual variability: not everybody should or does normatively suffer from mental illness.

Accommodating individual differences into the resource rationality framework, as it currently stands, is hindered by an emphasis on a modeling approach in which algorithmic hypothesis are developed by purely resource-rational considerations, and experiments that test group-level predictions of those hypothesis. This approach is at odds with characterizing individual differences. Indeed, Lieder and Griffiths explicitly argue against the alternative approach, common in computational cognitive neuroscience, of fitting components of models to human behavior. This latter approach could provide parametric accounts of individual differences in terms of component processes.

Nevertheless, even the current framework provides for some inter-individual variability. First, though viewed as a constant in the target article, individuals certainly differ in their capacity for, and hence cost of, cognitive operations. For one, some costs will be sensitive to representation and hence depend on individual experience. Individual cost functions might be measured through behavioral tasks while assuming resource-rational optimality with respect to that individual cost function. Such cost estimates could possibly assist in the development of tailored cognitive interventions. As a concrete example, multiple lines of evidence link the representation of time-based costs to an operation of specific catecholaminergic neuromodulators (Constantino et al. 2017; Hauser et al. 2018). This implies that dysregulated cognitive processing caused by irregular time-based costing could be amenable to pharmacological modulation.

Second, the tradeoff between cost and utility must be optimized with respect to an individual's environment. Similar to how different moods might reflect adaptive cognitive actions for different environmental states, inter-individual differences in cognitive heuristics may reflect adaptation to different environments. Mental illness may reflect cognitive strategies that are adaptive in certain environments. Though difficult, it is tantalizing to consider the possibility of using the resource rationality framework to characterize the environments that an individual's cognitive strategy might be optimized for.

Such considerations can also motivate research into the meta-learning processes by which individuals arrive at resource-optimal cognitive strategies. Aspects of cognitive behavioral therapy aimed at "cognitive restructuring" attempt to change an individual's cognitive heuristics so as to make them more adaptive (Beck 1979). For example, in cognitive therapy patients are asked to identify a recent emotion triggering situation (e.g. "I wasn't invited to a party"), record the beliefs they had in that situation (e.g. "everyone hates me"), and then critically evaluate this belief (e.g. "how would someone else think about this?)". Through the lens of resource-rational analysis, such approaches can be conceptualized as individuals learning to perform cognitive actions that lead to more adaptive inference regarding their current situations. More recent metacognitive treatments (Wells 2011) could even be viewed as altering the very cost function by which internal actions are evaluated.

In conclusion, the notion of resource rationality raises tantalizing possibilities in the realm of mental illness, and we encourage an expanded methodological approach that embraces individual differences.

Sampling as a resource-rational constraint

Adam N. Sanborn[a], Jianqiao Zhu[a], Jake Spicer[a] and Nick Chater[b]

[a]Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom and [b]Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom.
a.n.sanborn@warwick.ac.uk
J.Zhu@warwick.ac.uk
j.spicer@warwick.ac.uk
Nick.Chater@wbs.ac.uk
https://warwick.ac.uk/fac/sci/psych/people/asanborn/
https://warwick.ac.uk/fac/sci/psych/people/zjianqiao/
https://www.wbs.ac.uk/about/person/nick-chater/

doi:10.1017/S0140525X19001584, e22

Abstract

Resource rationality is useful for choosing between models with the same cognitive constraints but cannot settle fundamental disagreements about what those constraints are. We argue that sampling is an especially compelling constraint, as optimizing accumulation of evidence or hypotheses minimizes the cost of time, and there are well-established models for doing so which have had tremendous success explaining human behavior.

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

In the target article, the case for resource-rational analyses is made in general terms: It is a widely-applicable method for identifying how to best use cognitive resources given a set of cognitive constraints, and the long list of successes of this approach shows how resource-rational analyses explain a wide range of behavior. We are sympathetic to the overall thrust of the article, and particularly the argument that resource-rational analyses are useful for choosing between models with common cognitive constraints. Resource rationality provides a principled method for identifying how cognitive resources are used to solve tasks while assisting in identifying the important cognitive constraints.

But a key challenge for resource-rational analyses, which was highlighted in the target article, is identifying what the key cognitive constraints are. The long list of success in the target article is a heterogeneous one – it is comprised of many different approaches that are responding to different cognitive constraints, including neural constraints, representational constraints, time constraints, and attentional constraints, among others.

Researchers have tended to focus on a single constraint, rather than looking at them jointly. And indeed, different constraints do not necessarily all sit comfortably with one another, nor are they jointly necessary to explain behavioral biases. For concreteness, we focus on one of the topics discussed in the target article: biases in human probability judgments (Tversky & Kahneman 1974).

Several explanations have been advanced for these biases which appeal to resource rationality as a justification. One of the most influential explanations is that these biases are the result of estimating the probability of complex events (i.e., conjunctions and disjunctions of events) by averaging individual event probabilities together, rather than combining them correctly (Fantino et al. 1997). A resource-rational justification for averaging is that it is more accurate in the presence of internal or external noise than the correct combination rule (Juslin et al. 2009).

Models based on quantum probability have also been used to explain these behavioral biases, and make predictions that are similar to those of averaging. However, the underlying mechanisms of these models are very different from averaging, and also have a different resource-rational justification: instead of appealing to robustness to noise, they are justified as conserving representational resources (Busemeyer et al. 2011).

The third approach is covered in the target article: that people do not have access to their subjective probabilities, but are able to generate samples of events from either memory or an internal model. After an infinite number of samples, people could in principle recover their subjective probabilities exactly; but sampling is slow and effortful. With small samples, biases are introduced according to where sampling begins and by how small samples are converted into estimates. The resource-rational justification here is that generating samples takes time and effort – people make judgments and decisions with a small number of samples to optimally allocate time between different opportunities and challenges (Dasgupta et al. 2017; Sanborn & Chater 2016; Zhu et al. 2018a).

These three explanations appeal to very different, and likely mutually exclusive, cognitive constraints. As a result, resource rationality cannot be used to directly adjudicate between them. The best way to do so remains designing clever experiments for which the candidate models make different predictions. However, we argue that because resource rationality is part of the argument for each of these explanations, it is still useful to evaluate how compelling the cognitive constraints are and how well resource rationality has been applied.

We believe the cognitive constraint of sampling (in a broad sense, e.g., generating evidence or hypotheses in proportion to underlying probabilities) is especially compelling, as it is well-established both theoretically and empirically. In many contexts, the sampling process is slow and serial (Maylor et al. 2001), and therefore it is clearly important to optimize this time cost. Resource rationality is a starting point for many models using sampling: How to optimally stop sampling is well established, both for accumulating until a target confidence is reached and for stopping as soon as the expected cost exceeds the expected gain (Bogacz et al. 2006; Vul et al. 2014; Wald 1950). Models based on sequential sampling and optimal stopping have been extremely successful in both perceptual decision-making, and in wider forms of decision making (Ratcliff & McKoon 2008; Shadlen & Shohamy 2016). Indeed, sampling limitations underlie other examples discussed in the target article: Why people probability match, and why very good and very bad events are over-weighted.

Other constraints, such as representational or process noise constraints, are less well-attested and their consequences less clear cut. For example, applying representational constraints require first establishing what the representations are, and the nature of cognitive representations is often controversial (Spicer & Sanborn 2019). Internal noise is commonly used as a constraint – and indeed individual neurons are noisy – but in aggregate this noise may be less important than it seems (Beck et al. 2012), and its consequences again depend on the form of the representation. Although some aspects of the sampling process do also depend on the representation (Dasgupta et al. 2017; Zhu et al. 2018b), the fundamental goal of minimizing the number of samples remains.

Finally, beyond its usefulness as a cognitive constraint, sampling also satisfies other desiderata of the resource-rationality approach. As resource-rational analyses start from formulating a computational solution to a problem, sampling from the posterior is a useful algorithmic constraint to consider, because samplers are general algorithms for performing inference. Sampling models also have a clear connection to artificial intelligence and statistics, where these methods are widely used in Bayesian inference, and as a result can ease transfer of knowledge between these fields and the cognitive and brain sciences. For these reasons and those above, sampling is a very compelling cognitive constraint for resource rationality to target.

The evolutionary foundations of resource-rational analysis

Armin W. Schulz

Department of Philosophy, University of Kansas, Lawrence, KS 66045.
awschulz@ku.edu
http://people.ku.edu/~a382s825/

doi:10.1017/S0140525X1900164X, e23

Abstract

Resource-rational analysis would profit from being integrated more explicitly with an evolutionary psychological perspective. In particular, by taking more strongly into consideration the fact that efficiency considerations are a key driver of the evolution of human and animal minds, it becomes clearer: (1) why it is reasonable to assume that cognitive mechanisms trade-off accuracy against effort, (2) how this trade-off occurs, and (3) how to overcome some of the challenges of resource-rational analysis.

Lieder and Griffiths argue that cognitive modeling should proceed from the assumption that mental processes are optimizing, but in a cognitively efficient manner: "cognitive mechanisms [should be expected to] trade-off accuracy against effort in an adaptive, nearly optimal manner." Although there is much that speaks in favor of this kind of resource-rational analysis, the authors underemphasize the evolutionary foundations of resource-rational analysis. This is problematic, as emphasizing these evolutionary foundations allows resource-rational analysis to be expanded and strengthened. Specifically, by taking into consideration the fact that efficiency considerations are a key driver of the *evolution* of human and animal minds (Schulz 2018), it becomes clearer: (1) *why* it is reasonable to assume that cognitive mechanisms trade off accuracy against effort, (2) *how* this trade-off occurs, and (3) how to overcome some of the *challenges* of this framework. Consider these three points in turn.

First, without placing the appeal to evolutionary biology front and center, it is not clear why resource-rational analysis works. *Why* should it be presumed that humans evolved in such a way as to "trade off accuracy against effort in an adaptive, nearly optimal manner"? In particular, why didn't humans evolve so as to "satisfice" without ever focusing on what's optimal (as has been suggested e.g., by Gigerenzer & Selten 2001)? To answer these questions, it is necessary to consider the biological costs and benefits of (a) accurate decision-making, and (b) time- and cognitive resource-hungry-decision making. That is, it needs to be specified why accurate decision-making is biologically advantageous, and why these biological advantages are tempered by the cognitive and temporal costs that come from accurate decisions. Why, exactly, is an organism's fitness increased by relying on accurate cognitive processes, and decreased by relying on ones that take much time, concentration, and attention? It is just not obvious why an organism's expected reproductive success is affected by the accuracy and efficiency of cognitive processes. A step in the direction of an answer lies in the fact that it is reasonable to see cognitive efficiency and neurobiological efficiency as correlated, so that resource-rational decision-making allows organisms to save energy in maintaining and updating their central nervous system – a fact that is evolutionarily important (see Schulz 2018, for more on this). In general, it is only by providing a detailed account of the biological costs and benefits of resource-rational decision-making that Lieder and Griffiths's framework can be put on a plausible foundation and be properly distinguished from rival frameworks like that of Gigerenzer et al.

Second, a closer look at the evolutionary pressures on cognitive processing allows for improved predictions about *when* the tradeoff between accuracy and efficiency is resolved in *which way*. A good illustration of this is the authors' point that "people often think only about which subgoal to pursue next and how to achieve it […]. This is suboptimal from the perspective of expected utility theory […]. The resource-rationality framework can reconcile this tension by pointing out that goal-directed planning affords many computational simplifications that make good decision-making tractable." This, however, leaves it open *when*, exactly, people should be expected to rely on subgoals in their decision-making – and when they should be expected to rely on a more general goal of maximizing their expected utilities (or some such). An appeal to considerations from evolutionary biology can help answer these questions. For example, for many mammals, helping offspring in need is a choice that often needs to be made quickly and which responds to a unique set of environmental variables (viz., that the organism in need is one's offspring). From an evolutionary biological perspective, it can thus be predicted that this choice will be driven by a separate subgoal for helping offspring in need (see also Piccinini & Schulz 2019). By contrast, the choices humans need to make in order to navigate through their complex social environments – such as who to marry or what kinds of coalitions to join – respond to a wide range of variables (who is related to who, who is in a collation with who, who has which kinds of social statuses, etc.), but can typically be made less efficiently (in a matter of days rather than seconds). From an evolutionary biological perspective, it can thus be predicted that *these* choices will *not* be driven by separate subgoals, but by a more general goal of doing the best one can in society (or some such) (see also Schulz 2018, Ch. 8). In this way, an evolutionary perspective is useful for making predictions about the details of resource-rational analysis that can then be tested further.

Third, an evolutionary perspective allows resource-rational analysis to overcome some of the challenges it faces. Lieder and Griffiths note that "people's performance during the process of adaptation to a new environment" is difficult to analyze with their framework. However, from an evolutionary perspective, an organism's ability to efficiently adapt to new environments is just as much under selection as their ability to make decisions in a given environment. For example, there are good evolutionary biological reasons for thinking that more computationally intensive cognitive processes have an easier time adjusting to new environments, but that they pay a price in terms of the time, concentration, attention, etc., they need for their execution (Schulz 2018). By taking an evolutionary perspective more explicitly into account, it would thus become possible to broaden resource-rational analysis so that it can be applied to questions about "people's performance during the process of adaptation to a new environment" after all.

For these reasons, a fully compelling resource-rational analysis would profit greatly from being integrated more tightly with a more explicit evolutionary psychological foundation. This should thus be the next step in the development of this otherwise compelling theoretical framework.

Representing utility and deploying the body

David Spurrett

Philosophy, University of KwaZulu-Natal, Durban 4041, South Africa.
spurrett@ukzn.ac.za
https://www.researchgate.net/profile/David_Spurrett

doi:10.1017/S0140525X19001602, e24

Abstract

Comprehensive accounts of resource-rational attempts to maximise utility shouldn't ignore the demands of constructing utility representations. This can be onerous when, as in humans, there are many rewarding modalities. Another thing best not ignored is the processing demands of making functional activity out of the many degrees of freedom of a body. The target article is almost silent on both.

The target article urges that our criteria of rationality shouldn't ignore resource limitations, indeed that, properly understood, the demands of effectively deploying limited computational resources provides a unifying basis for recent work on how humans and other animals deviate from traditional models of rationality. The argument is worryingly silent on two related problems – processing utility representations, and making action out of a body.

Traditional rational actor models, and the refinements discussed in the target article (Equations 1 to 4) include a utility function. Their appearance in accounts of rationality isn't surprising, because instrumentally understood rationality is a matter of effective pursuit of some goals. Influential arguments defend the view that an effective agent will, among other things, have goals that satisfy certain requirements of consistency (Ramsay 1931), and behave as if she assigned – and updated – subjective probabilities to current and future states of the world (including states consequent on her own actions), and selected actions that maximised expected utility (Savage 1954). Okasha places these lines of thinking in an evolutionary context, to argue that an agent that acts and chooses *as if* performing Bayesian updating is an optimal agent, and hence a plausible target, at least sometimes, for natural selection (Okasha 2013). These arguments are typically understood behaviourally. Unlike the target article they defend claims about what effective agents do, not how they work.

Clearly enough, though, one straightforward way to act *as if* having beliefs updated in certain ways, and a utility function with certain properties, is to *actually have* those beliefs and preferences. More specifically, and leaving representations of the world to one side, whether a cognitive architecture is rational or resource-rational, if it is going to try to maximise utility, it has to build and maintain representations that convert and integrate the different dimensions of cost and return that matter to the agent into a consistent measure of utility, or common currency.

This has implications for cognitive architecture: If a "biologically feasible mind" (which the target article explicitly aims at) is to perform operations involving utility, then it has, somehow, to generate states that represent the expected returns from actions or strategies, future world states, life history segments, etc. These complications are sometimes ignored in models of rationality, which fix utility by fiat ("consider an agent valuing one slice of pizza as much as two ice-cream cones") in order to focus – like the target article – on other technicalities. Most living agents, though, have to deal with a heterogeneous mixture of costs and returns. The costs include direct expenditure of energy, the depletion of specific "fuels" or resources such as water and salt, as well as time and exposure to various risks. The returns include hydration, nutrition, rest, access to mating opportunities, acquisition of nesting materials or control of a nesting site, and so forth. A plausible list of only the primary reinforcers in humans (Rolls 2013) enumerates almost 50, some of them – such as "hormones" and "facial expressions" – with many variations falling into them. Costs and returns in these reinforcers are clumped together in heterogeneous bundles out in the world. Even in ideal circumstances, where important facts can be cheaply and reliably detected, converting the many dimensions of cost and return into utility values is likely to be difficult and resource-hungry. Circumstances aren't generally ideal, in part because much of the living world consists of rivals and competitors rather than allies. In both appearance and behaviour plants and animals often take steps to conceal or misrepresent their identity and likely behaviour. Sterelny (2003) called this informational "hostility," and it means that merely tracking costs and returns will sometimes be subject to trade-offs between cost and accuracy involving a further kind of resource limitation.

Using utility to select actions also requires sensitivity to the demands of controlling the body. These complications are ignored in many models of rationality, which take functional actions as primitive. The target article itself doesn't mention the body, and it does, perhaps revealingly, describe the resource rational "brain B interacting with the environment" (caption to Figure 1). This isn't how it works at all. In fact the stock of actions of a big complex animal like a human depends on a large structured array of muscles and other effectors, which only produces functional activity when subject to appropriate and sometimes complicated patterns of activation and suppression. The components of the actions themselves have their own metabolic and opportunity costs. That is to say, the cognitive demands of action production (getting the array of capacities to do this rather than that, or to do anything functional at all) aren't *independent* of the problem of trying to maximise utility (Spurrett 2019).

Both of these problems can vary in their demands. The overheads of constructing utility representations increase with the number of types of rewards and costs to which the system is to be responsive, as well as the costs of detecting cues of them with acceptable accuracy. (Detecting an acceptable egg to brood can cost more if your species is exploited by cuckoos.) The overheads of constructing the actions that a body is mechanically capable of increase with the number of degrees of freedom available, and how many need to be coordinated to produce functional activity. Attempting to maximise utility in an agent with a real body, and a real suite of sensory transducers (external and internal) requires facing up to these trade-offs.

There's a choice to be made here: Either admit to developing an account of rationality tailored to agents with a telekinetic action repertoire and roughly magical power to detect predictors of utility, so that the brain really does interact with the environment, or take the body seriously. The body, as a source of various channels of information about the world, the thing whose needs are the dimensions out of which utility is made, and as the thing that has to be controlled to produce action, is home to additional important kinds of resource-constraints.

What is the purpose of cognition?

Aba Szollosi and Ben R. Newell

School of Psychology, University of New South Wales, Sydney 2052, Australia.
aba.szollosi@gmail.com
ben.newell@unsw.edu.au

doi:10.1017/S0140525X19001626, e25

Abstract

The purpose of human cognition depends on the problem people try to solve. Defining the purpose is difficult, because people

seem capable of representing problems in an infinite number of ways. The way in which the function of cognition develops needs to be central to our theories.

Lieder and Griffiths argue that human cognition can be understood using much of the same framework that we would use to understand how a cash register works: in terms of the system's function and its resources (Marr 1982). But because people are different from cash registers, perhaps the best framework to understand one is not all that useful to understand the other. A major problem stems from the different ways in which these systems solve the *correspondence problem* – the problem of generating representations of the environment that correspond to the actual environment (Hammond 2000).

The cash register does not need to solve this problem – it has already been solved by its programmer. Because the programmer decided what the function of the machine will be, she presumably included all the necessary algorithms that can generate representations of the desired aspects of the world. For example, the machine needs an algorithm that generates representations of the value of items based on the key-presses of its user. Although such a heuristic will allow the cash register to represent this aspect of the environment well, it will also only ever represent these aspects.

Lieder and Griffiths suggest that evolution acted in a similar way to this programmer, and endowed people with algorithms that are optimised to generate representations of the environment. Thus, people do not need to solve the correspondence problem, because it has already been solved by evolution. According to this view, the difference between cash registers and people is quantitative: people have more algorithms – an "adaptive toolbox" of heuristics that generate environmental representations for them.

The shared underlying idea is that representation generation effectively consists of the matching of environmental input with already stored representations of potential environments (cf., Lieder & Griffiths 2017). For the cash register, this amounts to, for example, the deterministic matching of the key-press with the representation of the corresponding number. For people, the inputs are environmental features that were deemed relevant by evolution, some summary of which are then matched to the representation of the most similar problem. The main aim of resource-rational analysis is to find the solution to this problem with the highest expected value while also taking computational costs into account. This approach presupposes a clearly definable function for the system (i.e., a fixed environmental representation), which is defined by the researcher (target article, Box 2, Step 1).

But what if human cognition is not just a bag of tricks? It is not difficult to conceive how evolution would have favoured a cognitive system that can represent the environment more flexibly. Humans seem to be capable of representing any possible environment (and even impossible ones) and use that knowledge to learn and make decisions – in other words, representation generation in humans seems to be universal (cf., Deutsch 2011). Such a view of human cognition suggests that people are qualitatively different from the cash register in the sense that they can actually try to solve the correspondence problem. In fact, it suggests that people are more similar to scientists than to cash registers; and just like scientists, people need to generate and select the best out of many possible and plausible representations.

The flexibility in how people generate representations is demonstrated both by observations that their representations of the *same* environment can show considerable differences (e.g., Gaissmaier & Schooler 2008; Schulze & Newell 2016), and by observations that such representations can be improved on (e.g., Szollosi et al. 2019). The potential explanations that (a) these are in fact not new or improved representations, but result from the misapplication of strategies that evolved in a different evolutionary milieu; or (b) that people assess all potentially relevant features of the environment are both unsatisfactory. The former increases the flexibility of the model to an extent that there is almost nothing it cannot account for. The latter is computationally impossible.

This is not to say that people do not rely on heuristics to generate representations of the environment. Our argument is only that we should appreciate the flexibility of this generation process in our models, instead of substituting it with fixed representations based on flexible assumptions of the researcher. A promising way for such investigations would be the use of simple yet diagnostic manipulations of purportedly relevant features of the environment followed by thorough probing of people's knowledge about these features (e.g., Tran et al. 2017; also see, Newell & Shanks 2014).

The usefulness of resource-rational analysis hinges on the assumed similarity in representation generation between cash registers and people, because only under such conditions can the purpose of the system be clearly defined. We argued against this assumption: people seem capable of generating representations of anything, whereas cash registers can only represent things that they were programmed to represent. This difference leads to a radically different computational level question about cognition. Instead of asking what the purpose of people's cognitive mechanisms are in terms of prototypical evolutionary or learning environments, we could ask what purpose they serve in achieving the universality of representation generation. Such universality makes characterising the function of human cognition elusive, because with every new representation of a problem, the person generates a new function. Theories of human cognition need to clarify the development of purpose not merely presume its existence.

Beginning with biology: "Aspects of cognition" exist in the service of the brain's overall function as a resource-regulator

Jordan E. Theriault[a], Matt Coleman[a], Mallory J. Feldman[b], Joseph D. Fridman[a], Eli Sennesh[c], Lisa Feldman Barrett[a,d,e,1] and Karen S. Quigley[a,f,1]

[a]Department of Psychology, Northeastern University, Boston, MA, 02115; [b]Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599; [c]Khoury College of Computer Sciences, Northeastern University, Boston, MA, 02115; [d]Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, MA, 02129; [e]Department of Psychiatry, Massachusetts General

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Hospital and Harvard Medical School, Charlestown, MA, 02129 and [f]Department of Veterans Affairs, Edith Nourse Rogers Memorial (VA) Medical Center, Bedford, MA, 01730.
jordan_theriault@northeastern.edu, http://www.jordan-theriault.com/
m.coleman@northeastern.edu
mjfeld@live.unc.edu, https://malloryjfeldman.com
j.fridman@northeastern.edu, http://josephfridman.com/
sennesh.e@husky.neu.edu, https://esennesh.github.io/
l.barrett@northeastern.edu, https://www.affective-science.org/
K.Quigley@northeastern.edu, https://www.affective-science.org/

doi:10.1017/S0140525X19001705, e26

Abstract

Lieder and Griffiths rightly urge that computational cognitive models be constrained by resource usage, but they should go further. The brain's primary function is to regulate resource usage. As a consequence, resource usage should not simply select among algorithmic models of "aspects of cognition." Rather, "aspects of cognition" should be understood as existing in the service of resource management.

In their target article, Lieder and Griffiths suggest that efficient use of resources (e.g., psychological, neurobiological, and metabolic) must play a central role in computational models of cognition. We strongly agree, but in our view, the authors could go even further. Lieder and Griffiths suggest that resource-rational models should first identify "an aspect of cognition, formulated as a problem and its solution," and subsequently select an algorithm that "optimally trades-off resources and approximation accuracy" (Box 2). We suggest that "aspects of cognition" are not productively studied in isolation. Instead, "aspects of cognition" are best understood as a means to the functional end that brains evolved to serve: regulating the distribution and delivery of resources throughout the body – a balancing act called allostasis (Sterling 2004; 2012; Sterling & Eyer 1988; Sterling & Laughlin 2015). Resource usage, then, should not simply adjudicate among algorithms; rather, regulating resource usage is the brain's overall computational goal.

Centering computational modeling around "aspects of cognition" propagates an assumption that "aspects of cognition" can be separated and studied in isolation (Barrett 2019). In the nineteenth century, when psychology emerged from the tradition of mental philosophy as independent science, it inherited a list of mental faculties, or "aspects of cognition." Such "aspects of cognition" include categories such as attention, memory, and language, and in parallel, "aspects of emotion" with categories such as fear, anger, and sadness. An aspect of emotion like "fear" is thought to arise when dedicated survival-relevant mechanisms are implemented (Cook & Mineka 1989; Fanselow 2018; Fanselow & Lester 1988; Mineka et al. 1984); yet, researchers have generally failed to identify any consistent neural architecture implementing these mechanisms (Barrett 2016; Guillory & Bujarski 2014; Westermann et al. 2007). If the goal was only to solve a computational problem (e.g., how best to store and retrieve memories in a computer) then isolating "aspects of cognition" would be no issue. However, resource-rational analysis is ambitious *precisely because* it aims to do more: it aims to leverage resource limitations to better model how human brains solve problems. Considering "aspects of cognition" separately (e.g., language separate from memory; memory separate from attention; and attention separate from perception) is counterproductive, as the promise of resource-rational analysis is that it might allow these "aspects of cognition" to be compared on the common currency of resource usage (ideally, a measurable resource; e.g., metabolic costs of neuronal signaling; Attwell & Laughlin 2001; Niven & Laughlin 2008; Zénon et al. 2019). "Aspects of cognition" must be recognized as parts of a whole, and modeled in the context of the brain's general function within organisms.

Research in neuroanatomy, neurophysiology, and signal processing converges on the hypothesis that the brain's general function is to regulate metabolism and energy balance within an organism (Barrett & Finlay 2018; Chanes & Barrett 2016; Kleckner et al. 2017; Sterling 2012; Sterling & Laughlin 2015). Regulating energy balance is difficult: it requires that an organism uses "aspects of cognition" to find and extract resources from its environment, that an organism redirects and trade-offs resource usage within itself, and that an organism predicts what resources are needed to keep itself alive. A brain implements these processes and behaviors. An organism's brain uses perception to identify and navigate toward resources in the environment (while avoiding predators who are hunting for resources also); an organism's brain uses memory to revisit resource-rich locations and avoid resource-poor ones; and an organism's brain uses attention to prioritize sensory signals that are relevant to survival. A brain manages energy tradeoffs within the various systems of its body – for example, during peak effort, cardiac output is diverted away from some organs (e.g., liver and kidneys) to others (e.g., heart, lungs, skeletomotor muscles; Sterling & Laughlin 2015; Weibel 2000), investing energy in pressing concerns (e.g., escaping a predator) and divesting energy from concerns that can wait (e.g., digestion, immunoregulation). Finally, a brain manages resources efficiently by predicting what resources will be needed later and preparing to satisfy those needs before disruptions arise – for example, an organism should be motivated to drink (or seek water) before internal supplies are exhausted (Sterling 2012). These predictions require a major investment: the brain must continually run an internal model of the organism's body in its niche (i.e., the sensory aspects of the world that are relevant to its survival; Barrett 2016; Chanes & Barrett 2016; Clark, 2013; 2015; Denève & Jardri, 2016; Friston, 2010; Friston et al. 2017; Hutchinson & Barrett 2019; Seth 2015; Shadmehr et al. 2010). However, *even this predictive model is cost-effective*: by issuing predictions, and by encoding only prediction error, organisms can limit resources spent on neuronal signaling (Sengupta et al. 2013; Theriault et al. 2019; Zénon et al. 2019), encoding only signals that the predictive model did not already anticipate (i.e., prediction error; Chanes & Barrett 2016; Hutchinson & Barrett 2019; Shannon & Weaver 1949/1964; Theriault et al. 2019).

"At its biological core, life is a game of turning energy into offspring" (Pontzer 2015), meaning that if Lieder and Griffiths wish to model how a brain works, then resource usage must be a central (if not *the* central) computational concern. Brains did not evolve for animals to think or see or feel – they think, see, and feel because doing so regulates a body with resource-hungry systems. A resource-rational approach can set researchers in the right direction – constraining computational models using a common currency of resource usage – but the journey will be long unless computational goals are formulated using the empirical trail blazed by biology and neuroscience.

Note

[1] These authors jointly supervised this work.

Authors' Response

Advancing rational analysis to the algorithmic level

Falk Lieder[a] and Thomas L. Griffiths[b]

[a]Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany and [b]Departments of Psychology and Computer Science, Princeton University, Princeton, New Jersey 08544, USA.
falk.lieder@tuebingen.mpg.de; https://re.is.mpg.de
tomg@princeton.edu; https://psych.princeton.edu/person/tom-griffiths

doi:10.1017/S0140525X19002012, e27

Abstract

The commentaries raised questions about normativity, human rationality, cognitive architectures, cognitive constraints, and the scope or resource rational analysis (RRA). We respond to these questions and clarify that RRA is a methodological advance that extends the scope of rational modeling to understanding cognitive processes, why they differ between people, why they change over time, and how they could be improved.

We appreciated the diverse range of views reflected in the comments and we now face the task of composing the best response we can produce given the constraints imposed by our deadline, the word limit, and our opportunity cost. So, we will try to put our theory into practice. The commentaries raised a wide range of questions, concerns, and suggestions. To address them efficiently, we have grouped them into six sections. In the first section, we apply ideas from resource rational analysis (RRA) to commentators' examples of human errors and argue that RRA is useful even if people are only roughly resource rational. In the second section, we respond to commentaries that were concerned with the normative status of resource rationality and the philosophical and evolutionary foundations of RRA. In the third section, we synthesize and discuss the commentators' proposals for augmenting RRA with limits on what can be postulated as a cognitive constraint. In the fourth section, we discuss that RRA can be applied to different types of cognitive architectures. In the fifth section, we synthesize and discuss the commentators' thoughts on how incorporating cognitive constraints into rational models can broaden the scope of phenomena to which they are applicable. In the sixth section, we discuss how RRA can be extended beyond the cognition of a single individual. We conclude with a summary and future directions.

R1. RRA is useful even if people are only roughly resource rational

Several commentators pointed to behavioral results, thought experiments, and anecdotes about human judgments and decisions that deviate from certain intuitions or models of optimality. An RRA would leverage these findings to refine the model of what people are trying to do or identify how and why their heuristics fall short of the optimal solution. In the example of the apparently anti-Bayesian size–weight illusion mentioned by **Mandelbaum, Won, Gross, & Firestone**, a rational analysis might hypothesize that what people do is to first estimate the volume and the density of the object(s) from noisy observations and then multiply those estimates to estimate the object's pass. As illustrated in Figure R1, the reasonable assumption that people incorporate some prior knowledge according to which densities usually lie between those of the three light boxes and that of the heaviest box and volumes usually lie in between those of the single box and the combined volume of the three boxes – provides a parsimonious explanation for the size–weight illusion. (Similar results can be produced if mass is also assumed to be noisily observed in addition to volume and density.) For Mandelbaum et al.'s second example of a seemingly anti-Bayesian inference, belief polarization, multiple rational analyses have already been published (e.g., Cook & Lewandowsky 2016; Jern et al. 2014).

Contrary to the misconstrued framing by **Davis and Marcus**, RRA is not a retreat in an imaginary battle about whether people are rational. Rather, RRA is a methodological advance from modeling the function of cognitive abilities to modeling the underlying cognitive mechanisms. It moves forward the research program that David Marr (1982) initiated to reach an integrated understanding of the brain in which theories of its functions (computational level of analysis) inform and are informed by models of the underlying cognitive mechanisms (algorithmic level of analysis) and their biophysical realization in neural circuits (implementation level). It is critical to understand that, as **Rahnev** pointed out, RRA's assumption of bounded optimality is not a substantive claim about the mind but a methodological device to efficiently search through the endless space of possible mechanisms. Contrary to what Davis and Marcus claim, rationalizing irrationalities is NOT the goal of RRA. Rahnev correctly noted that (i) even if people were resource rational, RRA would be unable to prove this is the case and that (ii) even the successes of RRA cannot prove the assumption of resource rationality correct, they only prove it useful. The latter is not an accidental design flaw in our methodology. Rather, it reflects the fact that proving people to be rational has never been the goal. Instead, the goal of rational analysis and RRA has always been to understand what the mind is trying to do, how it does that, and why it does it that way. We agree that people are not perfectly resource rational and may be far from rational in certain situations. But as Rahnev correctly noted, RRA is useful even if we are trying to characterize non-resource rational humans, or as George Box put it, "All models are wrong but some are useful." Furthermore, RRA is useful not only as a methodology for understanding the mind, but also as a guideline for how to improve it as our preliminary work on cognitive tutors illustrates (Lieder et al. 2019a).

As **Rahnev** and our target article point out, there are many methodological challenges to testing substantive claims about resource (ir)rationality that nobody has been able to overcome yet. Any conclusions about people's resource (ir)rationality – including those of **Davis and Marcus** – thus come with serious methodological caveats. Some of these caveats can be addressed in future work, and we agree that measuring the relevant constraints is a valuable step toward enabling more accurate estimates of the extent to which the brain is resource (ir)rational.

Davis and Marcus claim that "a serious version of the bounded rationality view must presume that the tradeoff between effort and decision-making is made optimally" and then dismiss this hypothesis by appealing to common sense. Both are problematic. The former is problematic because demanding that people always optimally tradeoff the quality of their decisions against their cost is an unattainably high ideal that is incompatible with

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

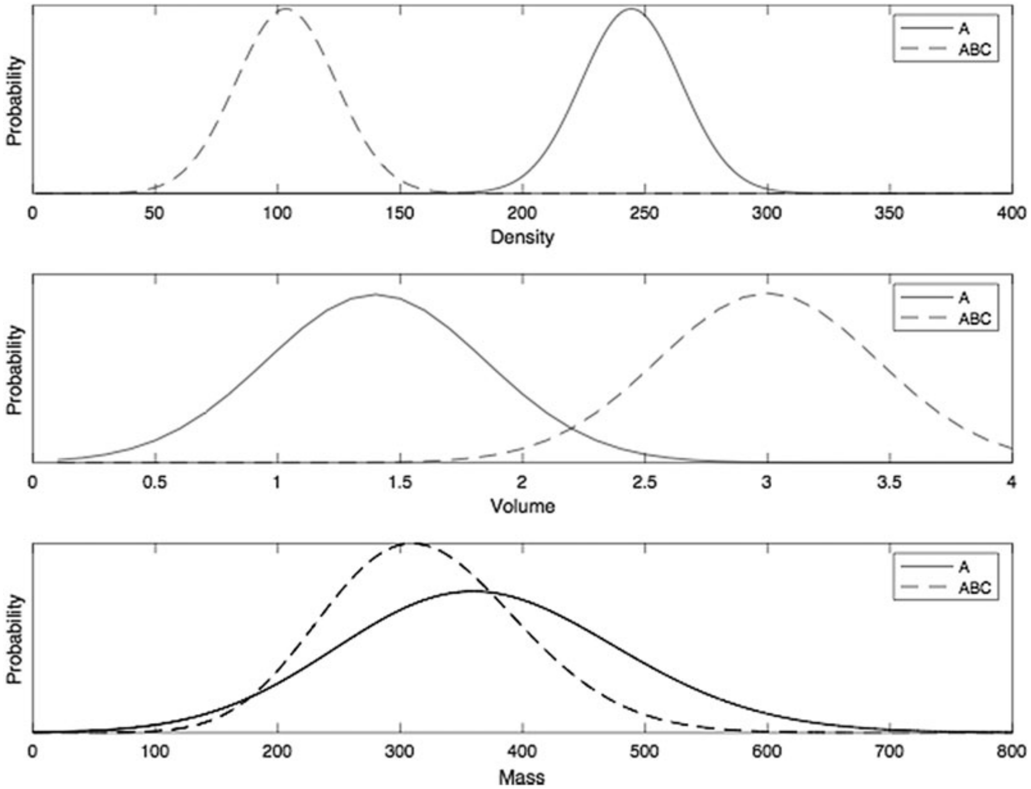


Figure R1. The “anti-Bayesian” perceptual illusion described by **Mandelbaum et al.** can be produced by a simple Bayesian model. This model infers the density and volume of an object based on noisy observations. The mass is then calculated from the inferred density and volume. If the prior favors larger volumes than the single container (A) and larger densities than the three containers (ABC) then the inferred mass will be higher for the single container. For simplicity, volume is measured in units of containers, density in grams per container.

the idea of resource rationality. Achieving this ideal would require optimal meta-decision-making which is even more computationally intractable than optimal decision-making itself. Resource rationality takes the cost of meta-decision-making into account and this means that even a resource rational decision maker will sometimes be too impulsive and other times overthink their decision. The resource rational framework makes it possible to derive precise predictions about the circumstances under which we should expect to see impulsivity and the circumstances in which we can expect to see overthinking. The underlying principle is that the mechanism by which the brain allocates control between different decision systems should make optimal use of the meta-decision system’s limited computational resources *on average* across all of the situations that the decision maker might encounter in the environment to which they are adapted. This means that it is not logically sound to refute resource rationality by pointing to individual situations in which people appear to occasionally make suboptimal tradeoffs. The distinction between a cognitive mechanism’s expected performance across all situations a person might encounter in their life versus its performance in one particular situation (or a handful of particular situations) invalidates the logic of the arguments that Davis and Marcus based on occasional errors in particular situations and thereby invalidates their strong conclusions that “bounded optimality casts virtually no light on what is and is not easy” and that RRA “predicts very little of the texture of actual human decision-making.”

We agree with **Davis and Marcus** that their examples of failures of human memory are not inevitable consequences of neural capacity constraints alone. This is why (resource-)rational analyses of memory emphasize the ecological distribution of problems to which memory is adapted. Our memory mechanisms appear to be optimized for evolutionarily important problems, such as navigation and social interaction, at the expense of being less effective for evolutionary less important problems, such as memorizing 10-digit numbers – even though there are real-life situations in which being able to memorize a 10-digit number would have high utility. Within the framework of RRA, one could hypothesize that people’s memory mechanisms are boundedly optimal for evolutionary environments rather than the environment of the twenty-first century.

Finally, **Davis and Marcus**’s misconstrual of the goal of our target article as arguing that people are rational led them to falsely accuse us of confirmation bias, claiming that we selectively reviewed research that can be construed as evidence for the hypothesis that cognitive biases are rational consequences of bounded cognitive resources. The truth is that our goal was to synthesize recent methodological advances. Within the constraints of our word limit, we have tried to provide a comprehensive – and thus, fair and unbiased – survey of previous applications of RRAs regardless of their conclusions about human (ir)rationality. In doing so, we have included several RRAs that identified interesting deviations from resource rationality. A grain of truth in the charge of confirmation bias might be that each methodology is usually preferentially applied in those areas where it is most useful. So, it is possible that there is a sampling bias or a publication in the literature, in the sense that researchers publish RRAs primarily about phenomena that are

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

roughly resource rational but do not attempt RRAs of phenomena that seem hopelessly irrational or let unsuccessful RRAs disappear in their metaphorical file drawers. So far, we have seen no evidence of this, but it seems plausible that this would happen with any new methodology.

R2. On RRA's philosophical and evolutionary foundations

Several commentators raised questions and concerns about the role of normative considerations in RRA as well as their nature, justification, and compatibility with evolutionary theory.

Colombo wondered how exactly the normative status of resource rationality can be justified. As highlighted in Equation 2, the normative status of resource rationality originates from the normative status of expected utility theory. That is, resource rational minds are optimal because they maximize the agent's utility in the long run within the limits of what the agent is capable of. Starting from this principle, we derived the definition of resource rational heuristics for a known environment (Equation 3) and then extended it to account for limited information about the structure of the environment (Equation 4). The environments or distributions over environments assumed in RRA are the ecological and evolutionary environments to which the agent is adapted and the utility function is meant to encode the goals of the organism. For this reason, we regard resource rationality as a (non-standard) version of ecological rationality. We would welcome future work that evaluates resource rationality against the desiderata for theories of ecological rationality highlighted by Colombo. Colombo characterized Equation 4 as saying that "rational agents ought to act so as to maximize some sort of expected utility, taking into account the costs of computation, time pressures, and limitations in the processing of relevant information available in the environment." We would thus like to clarify two things. First, Equation 4 defines rational cognitive mechanisms rather than rational behavior. Second, the agent is not expected to perform a cost-benefit analysis weighing the benefits of better decisions against the computational cost required to arrive at them. It is merely expected to carry out a simple heuristic that may have been discovered by evolution, learnt from past experiences, or copied from other people. Later on, Colombo rightfully highlights the need to specify under which conditions deviations from using the heuristic that would be most effective in a particular situation can or cannot be attributed to resource-rational heuristics for choosing heuristics. This problem can be solved by extending resource rationality by the addition of bounded-optimal meta-decision-making.

Colombo and **Kalbach** also raised the question whether and under which conditions violations of resource rationality should be considered as errors or cognitive biases. Kalbach took issue with us redefining the concept of "cognitive bias" as a violation of resource rationality while simultaneously endorsing inductive biases as a necessity for good scientific inferences. We would like to clarify that despite the lexical similarity inductive biases and cognitive biases are very different concepts from different fields. Inductive biases are neutral in valence – some kind of bias is necessary to support learning – whereas cognitive biases are defined by their deviation from some normative standard. Colombo highlights that there are many different types of rationality, such as epistemic rationality versus practical rationality, and that agents differ widely in their goals and cognitive constraints. He concludes that one, therefore, cannot diagnose irrationality from an agent's deviations from any single normative standard. We agree that this plurality renders blanket statements about people's (ir)rationality rather meaningless. As a constructive alternative, we would like to propose that (ir)rationality should always be measured relative to the individual's goals, preferences, and cognitive constraints, and the structure of their environment(s). Furthermore, the resulting assessment should be carefully qualified by exactly which type of rationality is being assessed and under which assumptions. The RRA framework can be used to incorporate many of these desiderata by adjusting its utility function, the bounds and costs on cognitive resource, and the distribution over possible environments. From an ethical perspective, we think it is important that these desiderata are considered if resource rationality is to be used to measure a person's rationality for purposes similar to those that IQ tests and personality inventories are used for.

Kalbach appeared to object to resource rationality as a normative standard because he thought that the resource rational cognitive mechanism is always a simple heuristic. But this is simply not true, because the optimal amount of thinking strongly depends on the person's utility function. That is, when the person's utility function and the nature of the situation make accuracy sufficiently more important than time and a more deliberate strategy performs sufficiently better than its heuristic alternatives, then extensive deliberation would be resource rational. In that case, relying on a simple heuristic would be resource-irrational. Thus, for a person who values advancing the frontiers of human knowledge above everything else, investing thousands of hours into astrophysics can be completely resource rational, even though for a person who does not value this kind of knowledge at all, the resource rational way of thinking might lead to serious misconceptions about the nature of the universe.

Given how diverse and flexible notions of rationality are, we second **Colombo**'s recommendation that researchers who use resource rationality to revisit the debate about human rationality should be very clear and precise about exactly what norms they are testing people against and word their conclusions accordingly. Contrary to what **Davis and Marcus** might think, we have absolutely no interest in perpetuating pointless debates about rationality based on terminological confusions.

Kalbach expressed serious concerns about the role of normative considerations in the descriptive enterprise of understanding the mind as it is. In his view, RRA is predicated on the naturalistic fallacy because it conflates what is with what ought to be. We would like to clarify that RRA clearly distinguishes between the optimal solutions to the problems solved by cognitive systems (i.e., what they "ought" to do) versus the cognitive/neural mechanisms they employ to realize that function (i.e., what actually "is" happening in the brain). We regard them as qualitatively different kinds of questions with different answers. So we do NOT confuse what is with what ought to be (or vice versa). But we do follow the legacy of David Marr (1982) in making the methodological assumption that to understand what a cognitive system does it is useful to attribute a function to it. This is a purely methodological device rather than a theoretical assumption. We fully subscribe to Darwinian evolution and we agree that there is no physical reality to concepts such as "purpose" and "function," but we think that these concepts are nevertheless useful for understanding the mind, developing models, and making predictions. We regret that our use of these terms was confusing, and we would like to clarify that the seemingly teleological components of RRA are purely methodological. That is, we use methodological assumptions of bounded optimality as a heuristic for generating

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

hypotheses about the mind/brain and then test them empirically. As the studies we reviewed in our target article illustrate, this approach has been very useful so far.

Szollosi and Newell also challenge the methodology of ascribing functions to the mind's cognitive systems. Their concern seems to be that unlike a cash register, the human mind does not have a fixed function to solve a given problem in a given representation but can invent its own problems and choose its own representations. In our view, the mind's capacity to flexibly adapt its representations is a computational resource that it can employ to realize the functions it has evolved to fulfill. From this perspective, RRA could be used to understand how and why people construct mental representations in the way they do. To the extent that people learn to solve their self-defined problems efficiently, RRA could also be used to model how people solve the problems they invented for themselves as if they were evolutionarily-engrained functions.

Contrary to **Kalbach** and **Szollosi and Newell**, **Theriault, Young & Barrett** and **Schulz** see great value in starting from evolutionarily-engrained functions. According to Theriault et al., one of the main functions that the brain evolved to serve is to regulate the distribution and delivery of limited resources throughout the body. We agree that regulating resource usage is an important function of the brain and would be happy to see RRA being applied to understand how the brain realizes this function.

Haas and Klein and Theriault et al. advocate extending RRA from individual cognitive processes to the entire brain. **Haas and Klein** argue that this is necessary to accurately capture how resource constraints emerge from and are negotiated by the competition between multiple processes, networks, or systems over multiple timescales. We welcome their proposal for holistic RRA and are happy to note that ongoing work by Musslick et al. (2016; 2017) and Segev et al. (2018) has already begun to implement it. Theriault et al. argue that "aspects of cognition" must be recognized as parts of a whole, and modeled in the context of the brain's general function within organisms. We agree that a complete theory of any component of the mind must encompass the entire organism and its environment. But since understanding complex systems can be very challenging, we also think that it is methodologically useful to initially focus on one of the system's modules as if it was an independent sub-system with a function of its own. This may be why, so far, RRA has been primarily applied to sub-systems that have been identified and isolated in previous psychological research.

Although **Kalbach** appeared to view the explanatory principle of resource rationality to be incompatible with Darwinian evolution, Schulz and Haas and Klein argued that RRA can and should be grounded in the theory of evolution. **Schulz** argues that the strong correlation between cognitive and neurobiological and metabolic efficiency provides an evolutionary foundation for the role of resource constraints and costs in RRA, and Theriault et al. emphasize that making efficient use of limited resources is essential from an evolutionary perspective. Schulz's evolutionary perspective also addresses Kalbach's misconception that deliberation can never be resource rational. Schulz argues that the evolutionary selection for the ability to adapt to changing environments makes deliberate reasoning resource rational in situations that cannot be handled by evolved simple heuristics. We agree with Schulz's perspective and we are looking forward to future work that will enrich RRA with evolutionary theory and RRAs of how people adapt to changing environments.

Haas and Klein point to additional insights from the study of evolution can inform RRA: satisficing, path dependencies, and competition between evolving and existing neural systems. We think that the insight that what can evolve easily strongly depends on what has evolved already might be an especially useful addition to RRA that speaks to the generous inclusion of capacities that appeared early in evolution in the cognitive architecture to which RRA is applied. We agree that re-use and overlap of neural pathways are critical for understanding why the capacity of certain cognitive systems is more constrained than the capacity of others.

R3. Introducing constraints on constraints

Several commentators (**Bates, Sims, and Jacobs** (**Bates et al.**); **Dimov**; **Sanborn, Zhu, Spicer, and Chater** (**Sanborn et al.**); **Ma & Woodford**) have correctly pointed out that identifying resource limitations is a critical bottleneck of RRA. Ma and Woodford pointed out there is currently no principled way to make those assumptions and that, consequently, extant RRAs differ widely in their assumptions about the nature of people's cognitive resources and their constraints. We agree with these commentators that this makes developing a principled methodology for identifying cognitive constraints an important direction for future work on RRA. Identifying cognitive constraints is challenging because any sub-optimality in performance could either result from a sub-optimal cognitive strategy, resource constraints, or a combination of both. We agree with Dimov that RRA itself can help us overcome this problem because the methodological assumption of bounded optimality solves the non-identifiability problem that usually arises when both the process and the cognitive architecture must be inferred at the same time.

Bates et al. proposed to require that all constraints must be formulated in terms of the information theoretic notion of channel capacity. We agree that channel capacity could provide a unifying language for modeling representational constraints. However, not all constraints are about representation. Some are also about how much computation can be performed on any given representation. **Dimov** proposed to ground assumptions about cognitive resources and their constraints in cognitive architectures such as ACT-R. We agree that this is a useful approach for leveraging the empirical findings that have already been built into these cognitive architectures, but there may be other computational resources and constraints that cognitive architectures do not capture yet. For instance, **Sanborn et al.** propose that one of those computational resources is sampling. Furthermore, **Ma and Woodford** point to RRAs where the relevant computational constraints are specified in terms of biophysical limits. We believe that different phenomena are best explained at different levels of analysis and/or abstraction. Furthermore, different cognitive systems (e.g., vision vs. relational reasoning) differ in their computational architectures and computational constraints. Thus, unlike Bates et al., we believe that there truly are different types of cognitive constraints. For instance, time constraints are conceptually different from limited working memory capacity. We therefore think that it makes sense that different RRAs emphasize different types of cognitive constraints.

Despite this, we do see great value in developing methodological principles for determining what the resource limitations are in a given domain at a given level of abstraction and to build bridges between the assumptions made at different levels of analysis. We

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

hope that our target article and the range of perspectives offered in the commentaries will help start an interdisciplinary conversation that will lead toward a unification of methodologies and a more principled approach to modeling cognitive constraints. Although we welcome **Bates et al.**'s idea to extended RRA with stronger constraints on what can be postulated as a constraint, it is not true that RRA does not have any constraints on constraints and runs the risk of overfitting and "just-so" theorizing. To the contrary, RRA already avoids overfitting and just-so stories by putting constraints on constraints; it does so by demanding that assumed constraints should be empirically grounded or empirically tested.

Ma and Woodford raised the question whether the sampling models from the one-and-done analysis, the resource rational anchoring-and-adjustment model, and the utility-weighted sampling model really optimize a linear combination of performance and resource cost. We can confirm that all three of these RRAs can be expressed in terms of Equation 3. In the case of the one-and-done analysis where the optimal number of samples is chosen so as to maximize expected performance minus the time cost of generating samples. In the resource rational anchoring-and-adjustment model, the number of adjustments is chosen so as to maximize expected reward minus the opportunity cost of time. The UWS model maximizes performance subject to a hard constraint on the number of samples. However, this is just a special case of Equation 3 in the target article where the cost of computation is constant across all heuristics.

We strongly agree with **Lewis and Howes** that RRA should be augmented with a principled theory guiding the modeler's assumptions about the utility function. Lewis and Howes argue that the utility function should reflect the agent's internal state (as in Equation 3) and view this as being inconsistent with the standard formulation of bounded optimality in Equation 2. To resolve this apparent inconsistency, we would like to clarify that although the utility function in Equation 2 scores the agent's entire life, the utility function in Equation 3 only scores its performance in making a single decision or judgment. One critical difference between these two settings is that the agent's belief state becomes worthless when the agent dies whereas the intermediate belief state following a single decision or judgment is valuable because it can inform future decisions. In our formulation, the value of the agent's belief states is grounded in the expected improvement in the value of world states brought about by its impact on future decisions. Thus, far from being unrelated or inconsistent, Equation 3 is a mathematical consequence of Equation 2. This reconciles Lewis and Howes's intuition that the agent's utility should depend on the agent's internal state with the original formulation of bounded optimality in Equation 2. Furthermore, we agree with Lewis and Howes that what we call the agent's belief state b in Equation 3 should be taken to include other aspects of the agent's internal state beyond its beliefs. Concretely, it should include all aspects of the agent's internal state that might impact its future decisions.

R4. Proposed computational architectures

RRA does not make a commitment to a particular computational architecture, specifying the terms of the tradeoff between utility and computational costs but not the kinds of computations or the way that those costs are denominated. Several commentaries proposed specific computational architectures, including sampling (**Sanborn et al.**), quantum computation (**Atmanspacher Basieva, Busemeyer, Khrennikov, Pothos, Shiffrin, and Wang** (**Atmanspacher et al.**); **Moreira, Fell, Dehdashti, Bruza, and Wichert** (**Moreira et al.**), and rule-based systems (**Dimov**).

The sampling approach advocated by **Sanborn et al.** is one to which we are very sympathetic, and it has been featured in many of our own resource rational models (e.g., Lieder et al. 2018a; 2018b). As they point out, the sampling approach is conducive to RRA. Since sampling is typically carried out sequentially, the cost of computation can be naturally formalized in terms of the opportunity cost of the time spent sampling. In addition to the reasons highlighted by Sanborn et al., sampling also gains psychological plausibility from the numerous natural psychological mechanisms that can instantiate it, including attending to the perceptual properties of an object (Gold & Shadlen 2007; Krajbich et al. 2012), retrieving experiences from memory in order to make a decision (Shadlen & Shohamy 2016), and mentally simulating the outcome of an interaction between physical objects (Battaglia et al. 2013).

Quantum computation provides an interesting alternative. As pointed out by **Atmanspacher et al.** and **Moreira et al.**, quantum probability takes a different approach to efficiently using resources, focusing on being able to capture a wide range of probabilistic outcomes without a significant increase in the representational resources required. However, we see this as presenting an alternative to traditional mechanisms of probabilistic computation rather than alternative to resource rationality itself. It is still possible to formulate resource rational models in the quantum framework. As **Sanborn et al.** point out, the relevant computational costs can be representational rather than algorithmic. Alternatively, we might imagine formulating resource rational problems of quantum computation, where the goal is to achieve the best possible result under a constraint on the number of quantum operations that can be performed (or equivalently, the size of a quantum circuit). As hinted at by Atmanspacher et al., the adoption of quantum probability may in itself be viewed as a solution to a problem of resource rationality: given the computational constraint that all computations need to be represented as operations on a vector space, quantum probability emerges as the appropriate way to perform probabilistic inference.

As **Dimov** points out and we discussed above, the identification of the computational architecture and corresponding costs is a challenging aspect of RRA. Dimov sees the solution as coming from the adoption of a universal cognitive architecture, reviving one of the classic goals of cognitive science. Historically, these cognitive architectures have focused on rule-based formalisms such as production systems to describe the generative capacity of human behavior, using chronometric analysis to link each of those computations with the time that it takes a human being to execute. We agree that given an architecture of this kind, RRA is particularly well-defined. The work of Lewis and Howes (Lewis et al. 2014) provides some compelling examples of the value of this approach. We agree with Dimov that the refinement of a unified theory of the mind should be pursued in parallel with RRA, with the two approaches being uniquely informative to one another.

R5. Considering constraints broadens the scope of rational models

We were encouraged by the wide range of applications that commentators envisaged for RRA. These applications include motor control (**Dounskaia & Shimansky**), psycholinguistics (**Dingemanse**), cognitive development (**Bejjanki & Aslin;**

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Persaud, Bass, Colantonio, Macias, and Bonawitz [**Persaud et al.**]), mental health (**Russek, Moran, McNamee, Reiter, Liu, Dolan, and Huys** [**Russek et al.**]), and even history (**Cowles & Kreiner**). Although we had not anticipated all the creative applications identified by the commentators, we did anticipate that integrating resource constraints into rational analysis would expand the scope of phenomena that it can explain. Accordingly, some of the application areas – specifically, cognitive development and mental health – did not come as a surprise. Although we do admit that history managed to sneak up on us.

In cognitive science, traditional rational models have up to three degrees of freedom: the prior, the data, and the utility function. But it has been thoroughly demonstrated that tweaking the utility function is not even enough to explain the variation of a single person's preferences within minutes (e.g., Allais 1953; Kahneman & Tversky 1979). Similarly, because the outcome of Bayesian inference is a direct result of what goes into it (the posterior probability of a hypothesis is directly proportional to the product of its prior probability and the likelihood reflecting the probability of the observed data), all inter-individual differences in beliefs and inferences would have to be explained as a consequence of variation in the priors of those agents or the data to which they were exposed. But in research areas where the goal is to explain variation, either across human lifetimes or as a result of mental illness, variation in priors, data, and utility functions may not be enough to capture these phenomena.

RRA adds two additional degrees of freedom: the computational resources available to an agent and their corresponding costs. These extra degrees of freedom are exactly the kind of thing that can be expected to vary across the lifespan or be influenced by mental illness. As a child grows up, the repertoire of computations available to them will expand, and the computational costs of particular operations may decrease as a consequence of practice or maturation. In addition, as **Persaud et al.** discuss, the goals of the child might change over time, and as pointed out by **Bejjanki and Aslin**, developmental resource constraints may themselves support more effective learning. In cases of mental illness, the availability of cognitive resources may be diminished and the computational costs of engaging in certain kinds of cognition may increase. **Russek et al.** highlight some concrete examples of cases where exactly such changes are known to happen in specific forms of mental illness.

We are also sympathetic to **Russek et al.**'s suggestion that other forms of mental illness, including mood disorders, might be best understood as systematic deviations from resource rationality. This suggests that uncovering deviations from resource rationality could be as useful for elucidating the cognitive distortions and aberrant processes that constitute specific mental illnesses as demonstrating deviations from classical notions of rationality has been for advancing our understanding of the heuristics and biases of healthy people (Tversky & Kahneman 1974). Once these systematic deviations from resource rational strategies have been identified, it will be especially interesting to understand how they were learned, how they can be unlearned, and how people can learn to think, learn, and decide according to more effective, near-resource rational strategies instead. To address these questions, we are currently developing models of metacognitive reinforcement learning (Krueger et al. 2017; Jain et al., under review).

Metacognitive learning of more resource rational cognitive strategies might be one of the primary effect mechanisms of effective psychotherapy, and we agree with **Russek et al.** that cognitive behavior therapy can be understood as teaching people more resource rational cognitive strategies. This is congruent with the view that the goal of cognitive therapy is to make people more rational (Baron et al. 1990). Taking this perspective one step further, we would like to suggest that resource rationality could even be used as a prescriptive principle to guide the development of more effective therapies. That is, RRA could be used to create a curriculum of adaptive cognitive strategies for healthy and resilient thinking, learning, and decision-making. Our current work on automatic strategy discovery (Callaway et al. 2018a; Gul et al. 2018; Lieder et al. 2017) and cognitive tutors teach people resource rational cognitive strategies (Lieder et al. 2019a) is a step in this direction.

We were intrigued by the suggestion from **Cowles and Kreiner** that resource rationality might have an equivalently valuable role to play for understanding history, but in retrospect, this application draws on the same principle as the applications to cognitive development and mental illness. In a historical context, the variation across individuals does not occur within a single human life, or in a snapshot of a society, but across societies over time. Again, the extra degree of freedom provided by considering the cognitive resources available to agents provides a way to engage with this variation. As Cowles and Kreiner point out, this provides the capacity to understand the decisions of historical agents and how they might differ from our contemporary intuitions because their cognitive tools were different and because their environments taxed their cognitive resources in a different way. We anticipate that a similarly fruitful analysis could be applied across contemporary cultures, extending the scope of resource rational models through space as well as time.

R6. Beyond individual cognition

Several commentaries observed that our focus in introducing resource rationality and in surveying related literature was on the cognitive states of individuals. This focus is consistent with the historical emphasis of cognitive psychology, from which many of the studies we summarized were drawn, but we do not view it as a fundamental limitation of the framework. In particular, the directions highlighted in the commentaries – recognizing that minds are embodied, that cognition interacts with emotion, and that individuals are part of societies – represent interesting frontiers for research on resource rationality.

Spurrett highlights the role that the physical body plays in specifying utility functions and imposing computational costs. We are sympathetic to this argument. One of the merits of resource rationality is that it provides a framework in which to explore the tradeoffs between these utilities and costs. While only implicit in the target article, we also view physical embodiment as playing an important role in defining the kinds of computational problems that human beings have to solve. For example, one significant constrained resource is being able to resolve visual information with high fidelity in only a small portion of the retina, turning the control of eye movements during decision-making into a problem that can be analyzed from the perspective of resource rationality. We view the problem of appropriately integrating biological constraints into resource rational models as an interesting direction for future research. Indeed, **Dounskaia and Shimansky** provide a nice example of such an approach.

Kauffman is concerned with the place of emotion in RRA. Rationality and emotion have long been held up as being at odds with one another. But there is also a tradition of pointing out the role that emotional responses can play in producing adaptive behavior, particularly in the context of interpersonal interaction (e.g., Frank 1988). Resource rationality provides a path to the resolution of this apparent contradiction, because the apparent antagonism between rationality versus emotion does not carry over into the resource rational framework. To the contrary, the computational efficiency of emotional mechanisms might make them resource rational in time-critical situations, and in certain situations, emotional mechanisms may be resource rational because they lead to better decisions than deliberation. Furthermore, emotions, such as anxiety, can guide the efficient allocation of cognitive resources to important problems, such as planning how to survive (Gagne et al. 2018).

We agree with **Russek et al.** that emotions can be understood in terms of resource-efficient computational mechanisms, but we would like to clarify being resource rational does not require solving the meta-decision-making problem optimally – instead, a resource rational agent would select computations by a boundedly optimal heuristic. Furthermore, RRA can also illuminate how emotions and cognition interact (Krueger & Griffiths 2018). An extensive body of work that has emphasized that there are at least three distinct decision systems: the instinctive Pavlovian system that is responsible for emotional biases, a deliberative system that supports effective goal pursuit through flexible reasoning, and a model-free reinforcement learning system that leads to inflexible habits (van der Meer et al. 2012). Exactly how those systems interact is an open problem that RRA could be used to solve. One proposal is that the model-based system generates simulated data – through a kind of introspection – that is then used to refine model-free learning (Gershman et al. 2014). Another proposal is that deliberation is used to refine the valuation of past experiences in the light of new information and to update the agent's habits accordingly (Krueger & Griffiths 2018). This latter perspective instantiates the idea that our emotions teach us how to become more resource rational by allowing our regrets to improve our computationally-efficient, habitual response tendencies.

Both **Ross** and **Dingemanse** point out that the formulation of RRA in the target article assumes an agent facing a problem that is generated by nature, while many of the problems that human beings have to solve require interacting with other agents. This creates a situation where the strategies adopted by one agent influence the environment experienced by another – a situation that is very familiar to any student of game theory. We do not foresee any fundamental obstacles to extending resource rationality to such situations. Indeed, we anticipate that this approach can be used to define models like those currently used in behavioral game theory (e.g., Camerer & Hua Ho 1999), but derived from the principle of optimization that underlies resource rationality. First steps in this direction have been taken by Halpern and Pass (2015). Beyond game theory, we agree with **Dingemanse** that language use represents a particularly rich territory for exploring this approach, including examining the extent to which speakers modify their linguistic choices based on assumptions about the cognitive load experienced by listeners.

R7. Summary and Conclusion

RRA is a new modeling paradigm that integrates the top-down approach that starts from the function of cognitive systems with the bottom-up approach that starts from insights into the mind's cognitive architecture and its constraints. Combining the strengths of both approaches makes RRA a promising methodology for reverse-engineering the mechanisms and representations of human cognition. RRA is an important step toward realizing David Marr's vision that theories formulated at different levels of analysis can inform and mutually constrain each other. RRA contributes to this vision by bringing insights about the function of cognitive systems (computational level) and empirical findings about the system's constraints (implementational level) to bear on models of cognitive mechanisms (algorithmic level of analysis). RRA accomplishes this in a principled way that uniquely specifies what the cognitive mechanism should be according to its function and the constraints of the available cognitive architecture. As Dimov noted, this addresses the fundamental non-identifiability problems that have been holding back progress on uncovering cognitive architectures and cognitive mechanisms for a long time. The commentaries revealed that RRA is even more broadly applicable than our target article suggested. We are looking forward to seeing RRA facilitating progress in fields ranging from cognitive development to history. We are especially excited to see RRA applied to understanding mental illness and improving people's mental health.

We appreciated the commentators' suggestions for future methodological developments, including the establishment of limits on the constraints that can be postulated by RRA and the integration of insights from extant cognitive architectures and evolutionary theory. RRA is a brand-new modeling paradigm that will undoubtedly mature and develop and the dialogue started by our target article will likely accelerate this process. The commentaries also gave us the opportunity to clarify the methodological nature of the teleological and optimality assumptions of RRA. We hope that this has made it clear that we are not arguing that the human mind is (resource) rational but offering a methodology for understanding the human mind's somewhat suboptimal cognitive systems in terms of their function, mechanisms, and representations.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Aerts, D., Gabora, L. & Sozzo, S. (2013) Concepts and their dynamics: A quantum–theoretic modeling of human thought. *Topics in Cognitive Science* **5**:737–72. [HA]

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society* **21**(4):503–46. [rFL]

Allers, R. & Minkoff, R. (1994) *The Lion King*. Walt Disney Pictures. [CJK]

Allport, D. A., Antonis, B. & Reynolds, P. (1972) On the division of attention: A disproof of the single channel hypothesis. *The Quarterly Journal of Experimental Psychology* **24** (2):225–35. doi:10.1080/00335557243000102. [aFL]

Anderson, J. R. (1978) Arguments concerning representations for mental imagery. *Psychological Review* **85**(4):249–77. doi:10.1037/0033-295X.85.4.249. [aFL]

Anderson, J. R. (1983) *The architecture of cognition*. Psychology Press. [CD]

Anderson, J. R. (1989) A rational analysis of human memory. In: *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, ed. H. L. Roediger & F. I. M. Craik, pp. 195–210. Lawrence Erlbaum Associates. [KP]

Anderson, J. R. (1990) *The adaptive character of thought*. Psychology Press. [aFL, CD]

Anderson, J. R. (1991) The adaptive nature of human categorization. *Psychological Review* **98**:409–29. [CD]

Anderson, J. R. (1996) ACT: A simple theory of complex cognition. *American Psychologist* **51**(4):355. [aFL]

Anderson, J. R. (2007) *How can the human mind occur in the physical universe?* Oxford University Press. [CD]

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C. & Qin, Y. (2004) An integrated theory of the mind. *Psychological Review* **111**(4):1036–60. doi:10.1037/0033-295X.111.4.1036. [aFL]

Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998) An integrated theory of list memory. *Journal of Memory and Language* **38**:341–80. [CD]

Anderson, J. R. & Milson, R. (1989) Human memory: An adaptive perspective. *Psychological Review* **96**(4):703–19. doi:10.1037/0033-295X.96.4.703. [aFL, CD]

Anderson, J. R. & Schooler, L. J. (1991) Reflections of the environment in memory. *Psychological Science* **2**(6):396–408. doi:10.1111/j.1467-9280.1991.tb00174.x. [aFL, CD]

Anderson, J. R., Zhang, Q., Borst, J. P. & Walsh, M. M. (2016) The discovery of processing stages: Extension of Sternberg's method. *Psychological Review* **123**:481–509. [CD]

Anderson, M. L. (2010) Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* **33**(4):245–66. [JH]

Ariely, D. (2009) *Predictably irrational*. Harper Collins. [aFL]

Atkinson, R. C., Holmgren, J. E. & Juola, J. F. (1969) Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics* **6**(6):321–26. doi:10.3758/BF03212784. [aFL]

Atmanspacher, H. & Römer, H. (2012) Order effects in sequential measurements of non-commuting psychological observables. *Journal of Mathematical Psychology* **56**(4):274–80. [HA]

Attwell, D. & Laughlin, S. B. (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism* **21**(10):1133–45. Available at: https://doi.org/10.1097/00004647-200110000-00001. [JET]

Austerweil, J. & Griffiths, T. (2011) Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science* **35**(3):499–526. doi:10.1111/j.1551-6709.2010.01161.x. [aFL, ESD]

Bacon, P.-L., Harb, J. & Precup, D. (2017) The option-critic architecture. In: *Proceedings from AAAI-17: The 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* (San Francisco, CA), pp. 1726–34. [aFL]

Barlow, H. B. (1961) Possible principles underlying the transformation of sensory messages. *Sensory Communication* **1**:217–34. [WJM]

Baron, J., Baron, J. H., Barber, J. P. & Nolen-Hoekseman, S. (1990) Rational thinking as a goal of therapy. *Journal of Cognitive Psychotherapy* **4**(3):293. [rFL]

Barrett, L. F. (2017a) *How emotions are made: The secret life of the brain*. Pan Macmillan. [JET]

Barrett, L. F. (2017b) The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* **12**(1):1–23. Available at: https://doi.org/10.1093/scan/nsw154. [JET]

Barrett, L. F. & Finlay, B. L. (2018) Concepts, goals and the control of survival-related behaviors. *Current Opinion in Behavioral Sciences* **24**:172–79. Available at: https://doi.org/10.1016/j.cobeha.2018.10.001. [JET]

Barrett, L. F., Quigley, K. S. & Hamilton, P. (2016) An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**(1708):20160011. Available at: https://doi.org/10.1098/rstb.2016.0011. [JET]

Barrett, L. F. & Satpute, A. B. (2019) Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters* **693**:9–18. Available at: https://doi.org/10.1016/j.neulet.2017.07.045. [JET]

Barutchu, A., Crewther, D. P. & Crewther, S. G. (2008) The race that precedes coactivation: Development of multisensory facilitation in children. *Developmental Science* **12** (3):464–73. doi:10.1111/j.1467-7687.2008.00782.x. [VRB]

Basieva, I., Cervantes, V. H., Dzhafarov, E. N. & Khrennikov, A. (2019) True contextuality beats direct influences in human decision making. *Journal of Experimental Psychology: General*. Online April 25, 2019. Available at: https://psycnet.apa.org/doiLanding?doi=10.1037%2Fxge0000585. [CM]

Basieva, I., Khrennikova, P., Pothos, E. M., Asano, M. & Khrennikov, A. (2018) Quantum-like model of subjective expected utility. *Journal of Mathematical Economics* **78**:150–62. [HA]

Bass, I., Shafto, P. & Bonawitz, E. (2018) That'll teach ‘em: How expectations about teaching styles may constrain inferences. In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (Madison, WI). Cognitive Science Society. [KP]

Bates, C. J. & Jacobs, R. A. (2019) Efficient data compression leads to categorical bias in perception and perceptual memory. In: *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, July 24–27, Montreal, Canada. [CJB]

Bates, C. J., Lerch, R. A., Sims, C. R. & Jacobs, R. A. (2019) Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision* **19**(2):11, 1–23. [CJB]

Bateson, M., Healy, S. D. & Hurly, T. A. (2002) Irrational choices in hummingbird foraging behaviour. *Animal Behaviour* **63**(3):587–96. [aFL]

Batson, C. D. (1975) Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology* **32**:176–84. [EM]

Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. (2013) Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences* **110**(45):18327–32. [rFL]

Bays, P. M. (2014) Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience* **34**(10):3632–45. [WJM]

Beck, A. T. (1979) *Cognitive therapy of depression*. Guilford Press. [EMR]

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. & Pouget, A. (2012) Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron* **74**(1):30–39. doi:10.1016/j.neuron.2012.03.016. [aFL, ANS]

Beer, R. D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* **4**(3):91–99. [aFL]

Bejjanki, V. R., Knill, D. C. & Aslin, R. N. (2016) Learning and inference using complex generative models in a spatial localization task. *Journal of Vision* **16**(5):9. doi:10.1167/16.5.9. [VRB]

Bejjanki, V. R., Randrup, E. R. & Aslin, R. N. (2019) Young children combine sensory cues with learned information in a statistically efficient manner: But task complexity matters. *Developmental Science* e12912. doi:10.1111/desc.12912. [VRB]

Bender, A. & Beller, S. (2014) Mangarevan invention of binary steps for easier calculation. *Proceedings of the National Academy of Sciences* **111**(4):1322–27. doi:10.1073/pnas.1309160110. [MD]

Berniker, M., Voss, M. & Kording, K. (2010) Learning priors for Bayesian computations in the nervous system. *PLoS One* **5**(9):e12686. doi:10.1371/journal.pone.0012686. [VRB]

Best, J. R. & Miller, P. H. (2010) A developmental perspective on executive function. *Child Development* **81**(6):1641–60. [VRB]

Bhui, R. & Gershman, S. J. (2017) Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review* **125**(6):985-1001. doi:10.1037/rev0000123. [aFL]

Bless, H., Schwarz, N. & Kemmelmeier, M. (1996) Mood and stereotyping: The impact of moods on the use of general knowledge structures. In: European review of social psychology, vol. 7, ed. M. Hewstone & W. Stroebe, pp. 63–93. Wiley. [KP]

Blume, L. E. & Easley, D. (1984) Rational expectations equilibrium: An alternative approach. *Journal of Economic Theory* **34**(1):116–29. [WJM]

Böckler, A., Knoblich, G. & Sebanz, N. (2010) Socializing cognition. In: *Towards a theory of thinking*, ed. B. Glatzeder, V. Goel & A. Müller, pp. 233–50. doi:10.1007/978-3-642-03129-8_16. Heidelberg. [MD]

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. (2006) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**(4):700–65. doi:10.1037/0033-295x.113.4.700. [aFL, ANS]

Borst, J. P. & Anderson, J. R. (2017) A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology* **76**:94–103. [CD]

Bossaerts, P. & Murawski, C. (2017) Computational complexity and human decision-making. *Trends in Cognitive Sciences* **21**(12):917–29. doi:10.1016/j.tics.2017.09.005. [aFL]

Bossaerts, P., Yadav, N. & Murawski, C. (2018) Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B* **374**(1766):20180138. [aFL]

Botvinick, M. (2008) Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences* **12**(5):201–08. doi:10.1016/j.tics.2008.02.009. [aFL]

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

Botvinick, M. M. & Cohen, J. D. (2014) The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* **38**(6):1249–85. [JH]

Botvinick, M., Weinstein, A., Solway, A. & Barto, A. (2015) Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences* **5**:71–77. [CJB]

Boureau, Y.-L., Sokol-Hessner, P. & Daw, N. D. (2015) Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences* **19**(11):700-10 doi:10.1016/j.tics.2015.08.013. [aFL]

Bowers, J. S. & Davis, C. J. (2012a) Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin* **138**:389–414. [MC]

Bowers, J. S. & Davis, C. J. (2012b) Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget. *Psychological Bulletin* **138**:423–26. [MC]

Braine, M. D. (1978) On the relation between the natural logic of reasoning and standard logic. *Psychological Review* **85**(1):1–21. doi:10.1037/0033-295X.85.1.1. [aFL]

Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. (2017) Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review* **124** (3):301–38. doi:10.1037/rev0000061. [aFL]

Bray, D. (2009) *Wetware. A computer in every living cell*. Yale University Press. [KPK]

Brayanov, J. B. & Smith, M. A. (2010) Bayesian and "anti-Bayesian" biases in sensory integration for action and perception in the size–weight illusion. *Journal of Neurophysiology* **103**:1518–31. [EM]

Briggs, R. A. (2017) Normative theories of rational choice: Expected utility. In: *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta, Metaphysics Research Lab. Stanford University. [MC]

Brown, P. (2012) *Through the eye of a needle: Wealth, the fall of Rome, and the making of Christianity in the West, 350–550 AD*. Princeton University Press. [HMC]

Brown, P. (2015) *The ransom of the soul: Afterlife and wealth in early western Christianity*. Harvard University Press. [HMC]

Brown, R. L. (2013) What evolvability really is. *The British Journal for the Philosophy of Science* **65**(3):549–72. [JH]

Bruton, M. & O'Dwyer, N. (2018) Synergies in coordination: A comprehensive overview of neural, computational, and behavioral approaches. *Journal of Neurophysiology* **120**:2761–74. [ND]

Bruza, P. D., Wang, Z. & Busemeyer, J. R. (2015) Quantum cognition: A new theoretical approach to psychology. *Trends in Cognitive Sciences* **19**(7):383–93. [HA]

Buckingham, G. (2014) Getting a grip on heaviness perception: A review of weight illusions and their probable causes. *Experimental Brain Research* **232**:1623–29. [EM]

Buckingham, G. & Goodale, M. A. (2013) When the predictive brain gets it really wrong. *Behavioral and Brain Sciences* **36**:208–09. [EM]

Budiu, R. & Anderson, J. R. (2004) Interpretation-based processing: A unified theory of semantic sentence comprehension. *Cognitive Science* **28**:1–44. [CD]

Busemeyer, J. & Bruza, P. (2012) *Quantum models for cognition and decision*. Cambridge University Press. [CM]

Busemeyer, J. R., Pothos, E. M., Franco, R. & Trueblood, J. S. (2011) A quantum theoretical explanation for probability judgment errors. *Psychological Review* **118**(2):193–218. [ANS]

Buss, D. M. (1995) Evolutionary psychology: A new paradigm for psychological science. *Psychological Inquiry* **6**(1):1–30. [aFL]

Butko, N. J. & Movellan, J. R. (2008) I-POMDP: An infomax model of eye movement. *In: Proceedings from ICDL 2008: 7th IEEE International Conference on Development and Learning* (Monterey, CA), pp. 139–44. doi:10.1109/DEVLRN.2008.4640819. [aFL]

Calcott, B. (2014) The creation and reuse of information in gene regulatory networks. *Philosophy of Science* **81**(5):879–90. [JH]

Calkins, S. D. & Bell, M. A. E. (2010) *Child development at the intersection of emotion and cognition*. American Psychological Association. [KP]

Callaway, F., Gul, S., Krueger, P.M., Griffiths, T.L., Lieder, F. (2018a) Learning to select computations. In: *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference*. [arFL]

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M. & Griffiths, T. L. (2018b) A resource-rational analysis of human planning. In: *Proceedings from 40th Annual Conference of the Cognitive Science Society*. Cognitive Science Society. [aFL]

Callaway, F., Gul, S., Krueger, P. M., Griffiths, T. L. & Lieder, F. (in preparation). Discovering rational heuristics for risky choice. [aFL]

Camerer, C. & Hua Ho, T. (1999) Experience-weighted attraction learning in normal form games. *Econometrica* **67**(4):827–74. [rFL]

Caplin, A. & Dean, M. (2015) Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* **105**(7):2183–203. doi:10.3386/w19876. [aFL]

Caplin, A., Dean, M. & Leahy, J. (2017) *Rationally inattentive behavior: Characterizing and Generalizing Shannon Entropy*. NBER Working Paper No. 23652. National Bureau of Economic Research. [aFL]

Caplin, A., Dean, M. & Martin, D. (2011) Search and satisficing. *American Economic Review* **101**(7):2899–922. doi:10.1257/aer.101.7.2899. [aFL]

Carlson, S. M., Zelazo, P. D. & Faja, S. (2013) Executive function. In: *The Oxford handbook of developmental psychology: Vol. 1. Body and mind*, ed. P. D. Zelazo, pp. 706–742. Oxford University Press. [VRB]

Carver, C. S. & Scheier, M. F. (2001) *On the self-regulation of behavior*. Cambridge University Press. [aFL]

Cervantes, V. H. & Dzhafarov, E. N. (2018) Snow Queen is evil and beautiful: Experimental evidence for probabilistic contextuality in human choices. *Decision* **5**:193–204. [CM]

Chambers, C., Sokhey, T., Gaebler-Spira, D. & Kording, K. P. (2018) The development of Bayesian integration in sensorimotor estimation. *Journal of Vision* **18**(12):8. [VRB]

Chanes, L. & Barrett, L. F. (2016) Redefining the role of limbic areas in cortical processing. *Trends in Cognitive Sciences* **20**(2):96–106. Available at: https://doi.org/10.1016/j.tics.2015.11.005. [JET]

Charpentier, A. (1891) Analyse experimentale: De quelques elements de la sensation de poids. [Experimental analysis: On some of the elements of sensations of weight]. *Archives de Physiologie Normale et Pathologique* **3**:122–35. [EM]

Chater, N. & Oaksford, M. (1999) Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences* **3**(2):57–65. doi:10.1016/S1364-6613(98)01273-X. [aFL]

Chater, N., Tenenbaum, J. B. & Yuille, A. (2006) Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* **10**(7):287–91. doi:10.1016/j.tics.2006.05.007. [aFL]

Chi, M. T. & Ceci, S. J. (1987) Content knowledge: Its role, representation, and restructuring in memory development. In: Advances in child development and behavior, vol. 20, ed. H. W. Reese, pp. 91–142. Academic Press. [KP]

Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press. [MD]

Clark, A. (1997) *Being there*. MIT Press. [DRo]

Clark, A. (1998) Magic words: How language augments human computation. In: *Language and thought: Interdisciplinary themes*, ed. P. Carruthers & J. Boucher, pp. 162–83. Cambridge University Press. [MD]

Clark, A. (2003) *Natural born cyborgs*. Oxford University Press. [DRo]

Clark, A. (2006) Material symbols. *Philosophical Psychology* **19**(3):291–307. doi:10.1080/09515080600689872. [MD]

Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**(3):1–24. Available at: https://doi.org/10.1017/S0140525X12000477. [JET]

Clark, A. (2015) *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. [JET]

Clark-Polner, E., Wager, T. D., Satpute, A. B. & Barrett, L. F. (2016) Neural fingerprinting: Meta-analysis, variation and the search for brain-based essences in the science of emotion. In: *The handbook of emotion*, 4th edition, ed. L. F. Barrett, M. Lewis & J. M. Haviland-Jones, pp. 146–65. Guilford Press. [JET]

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990) On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review* **97**(3):332–61. [JH]

Colombo, M. (2019) Learning and reasoning. In: *The Routledge handbook of the computational mind*, ed. M. Sprevak & M. Colombo, pp. 381–96. Routledge. [MC]

Colombo, M., Elkin, L. & Hartmann, S. (forthcoming) Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science* (first online 03 August 2018). Available at: https://doi.org/10.1093/bjps/axy059. [MC]

Colombo, M. & Hartmann, S. (2017) Bayesian cognitive science, unification, and explanation. *The British Journal for Philosophy of Science* **68**:451–84. [MC]

Colombo, M. & Seriès, P. (2012) Bayes in the brain. On Bayesian modelling in neuroscience. *The British Journal for Philosophy of Science* **63**:697–723. [MC]

Constantino, S. M., Dalrymple, J., Gilbert, R. W., Varanese, S., Di Rocco, A. & Daw, N. D. (2017) A neural mechanism for the opportunity cost of time. *bioRxiv* **173443**. Available at: http://doi.org/10.1101/173443. [EMR]

Cook, J. & Lewandowsky, S. (2016) Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science* **8**(1):160–79. [rFL]

Cook, M. & Mineka, S. (1989) Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *Journal of Abnormal Psychology* **98**(4):448–59. Available at: https://doi.org/10.1037/0021-843X.98.4.448. [JET]

Cooper, W. S. (2001) *The evolution of reason*. Cambridge University Press. [MC]

Crone, E. A., Bunge, S. A., Van Der Molen, M. W. & Ridderinkhof, K. R. (2006) Switching between tasks and responses: A developmental study. *Developmental Science* **9**(3):278–87. [KP]

Cubitt, T. S., Perez-Garcia, D. & Wolf, M. (2015) Undecidability of the spectral gap. *Nature* **528**:207. [CM]

Damasio, A. R. (1999) *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt. [KPK]

Darnton, R. (1984) *The great cat massacre and other episodes in French cultural history*. Vintage. [HMC]

Dasgupta, I., Schulz, E. & Gershman, S. J. (2017) Where do hypotheses come from? *Cognitive Psychology* **96**:1–25. doi:10.1016/j.cogpsych.2017.05.001. [aFL, ANS]

Dasgupta, I., Schulz, E., Goodman, N. D. & Gershman, S. J. (2018) Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition* **178**:67-81. doi:10.1016/j.cognition.2018.04.017. [aFL]

Davidson, M. C., Amso, D., Anderson, L. C. & Diamond, A. (2006) Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* **44**(11):2037–78. [KP, VRB]

Davies, P. (2019) *The Demon in the machine: How hidden webs of information are solving the mystery of life*. Allen Lane Pub. [KPK]

d'Avray, D. L. (2010) *Rationalities in history: A Weberian essay in comparison*. Cambridge University Press. [HMC]

Daw, N., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8**(12):1704–11. doi:10.1038/nn1560. [aFL]

Dawes, R. M. & Mulford, M. (1996) The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes* **65**(3):201–11. [aFL]

Dayan, P. & Abbott, L. F. (2001) Theoretical neuroscience: Computational and mathematical modeling of neural systems, *1st edition*. MIT Press. [aFL]

de Barros, J. A. & Oas, G. (2015) Some examples of contextuality in physics: Implications to quantum cognition. *arXiv* **1512**.00033. [CM]

de Barros, J. A. & Oas, G. (2016) Some examples of contextuality in physics: Implications to quantum cognition. In: *Contextuality from quantum physics to psychology*, ed. E. Dzhafarov, J. Jordan, R. Zhang & V. Cervantes, pp. 153–84. World Scientific. [CM]

de Barros, J. A. & Suppes, P. (2009) Quantum mechanics, interference and the brain. *Journal of Mathematical Psychology* **53**:306–313. [CM]

Denève, S. & Jardri, R. (2016) Circular inference: Mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences* **11**:40–48. Available at: https://doi.org/10.1016/j.cobeha.2016.04.001. [JET]

Deutsch, D. (2011) *The beginning of infinity*. The Penguin Press. [AS]

Dickhaut, J., Rustichini, A. & Smith, V. (2009) A neuroeconomic theory of the decision process. *Proceedings of the National Academy of Sciences* **106**(52):22145–50. doi:10.1073/pnas.0912500106. [aFL]

Diggins, J. P. (1978) *The bard of savagery: Thorstein Veblen and modern social theory*. Seabury Press. [HMC]

Dimov, C. M. & Link, D. (2017) Do people order cues by retrieval fluency when making probabilistic inferences? *Journal of Behavioral Decision Making* **30**:843–54. [CD]

Dingemanse, M. (2017) On the margins of language: Ideophones, interjections and dependencies in linguistic theory. In: *Dependencies in language*, ed. N. J. Enfield, pp. 195–202. doi:10.5281/zenodo.573781. Language Science Press. [MD]

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G. & Enfield, N. J. (2015) Universal principles in the repair of communication problems. *PLoS One* **10**(9): e0136100. doi:10.1371/journal.pone.0136100. [MD]

Dolan, R. & Dayan, P. (2013) Goals and habits in the brain. *Neuron* **80**(2):312–25. doi:10.1016/j.neuron.2013.09.007. [aFL]

Dounskaia, N. (2005) The internal model and the leading joint hypothesis: Implications for control of multi-joint movements. *Experimental Brain Research* **166**:1–16. [ND]

Dounskaia, N. (2010) Control of human limb movements: The leading joint hypothesis and its practical applications. *Exercise and Sport Sciences Reviews* **4**:201–08. [ND]

Dounskaia, N. & Shimansky, Y. (2016) Strategy of arm movement control is determined by minimization of neural effort for joint coordination. *Experimental Brain Research* **234**:1335–50. [ND]

Dukas, R., ed. (1998a) *Cognitive ecology: The evolutionary ecology of information processing and decision making*. University of Chicago Press. [aFL]

Dukas, R. (1998b) Constraints on information processing and their effects on behavior. In: *Cognitive ecology: The evolutionary ecology of information processing and decision making*, ed. R. Dukas. University of Chicago Press. [aFL]

Dukas, R. (2004) Evolutionary biology of animal cognition. *Annual Review of Ecology, Evolution, and Systematics* **35**:347–74. [aFL]

Dzhafarov, E. N. & Kujala, J. V. (2014) Contextuality is about identity of random variables. *Physica Scripta* **T163**:014009. [CM]

Dzhafarov, E. N. & Kujala, J. V. (2016) Context-content systems of random variables: The contextuality-by-default theory. *Journal of Mathematical Psychology* **74**:11–33. [CM]

Eckstein, M. P. (1998) The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science* **9**(2):111–18. doi:10.1111/1467-9280.00020. [aFL]

Edwards, W. (1954) The theory of decision making. *Psychological Bulletin* **51**(4):380. [aFL]

Elliott, R., Zahn, R., Deakin, J. F. W. & Anderson, I. M. (2011) Affective cognition and its disruption in mood disorders. *Neuropsychopharmacology* **36**(1):153–82. Available at: http://doi.org/10.1038/npp.2010.77. [EMR]

Elman, J. L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* **48**(1):71–99. doi:10.1016/0010-0277(93)90058-4. [VRB]

Enfield, N. J. (2013) *Relationship thinking: Agency, enchrony, and human sociality*. Oxford University Press. [MD]

Enfield, N. J. (2017) *How we talk: The inner workings of conversation*. Basic Books. [MD]

Epley, N. & Gilovich, T. (2004) Are adjustments insufficient? *Personality and Social Psychology Bulletin* **30**(4):447–60. doi:10.1177/0146167203261889. [aFL]

Evans, J. S. B. (2003) In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences* **7**(10):454–59. [CM]

Evans, J. St. B. T. (2008) Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology* **59**:255–78. doi:10.1146/annurev.psych.59.103006.093629. [aFL]

Fanselow, M. S. (2018) Emotion, motivation and function. *Current Opinion in Behavioral Sciences* **19**:105–09. Available at: https://doi.org/10.1016/j.cobeha.2017.12.013. [JET]

Fanselow, M. S. & Lester, L. S. (1988) A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior. In: *Evolution and learning*, ed. R. C. Bolles & M. D. Beecher, pp. 185–212. Lawrence Erlbaum Associates, Inc. [JET]

Fantino, E., Kulik, J., Stolarz-Fantino, S. & Wright, W. (1997) The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review* **4**(1):96–101. [ANS]

Fawcett, T. W., Fallenstein, B., Higginson, A. D., Houston, A. I., Mallpress, D. E., Trimmer, P. C. & McNamara, J. M. (2014) The evolution of decision rules in complex environments. *Trends in Cognitive Sciences* **18**(3):153–61. [aFL]

Fechner, H. B., Pachur, T., Schooler, L. J., Mehlhorn, K., Battal, C., Volz, K. G. & Borst, J. P. (2016) Strategies for memory-based decision making: Modeling behavioral and neural signatures within a cognitive architecture. *Cognition* **157**:77–99. [CD]

Feng, S. F., Schwemmer, M., Gershman, S. J. & Cohen, J. D. (2014) Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience* **14**(1):129–46. doi:10.3758/s13415-013-0236-9. [aFL]

Festinger, L., Riecken, H. W. & Schachter, S. (1956) *When prophecy fails*. University of Minnesota Press. [EM]

Fischer, R. & Plessow, F. (2015) Efficient multitasking: Parallel versus serial processing of multiple tasks. *Frontiers in Psychology* **6**:1366. doi:10.3389/fpsyg.2015.01366. [aFL]

Fiser, J. & Aslin, R. N. (2002) Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America* **99**(24):15822–26. [VRB]

Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. (2010) Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences* **14** (3):119–30. doi:10.1016/j.tics.2010.01.003. [aFL]

Fitts, P. M. (1954) The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* **47**:381–91. [ND]

Fodor, J. A. (1987) Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In: *The robot's dilemma: The frame problem in artificial intelligence*, ed. Z. W. Pylyshyn, pp. 139–50. Ablex. [aFL]

Fox Tree, J. E. (1995) The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* **34**(6):709–38. doi:10.1006/jmla.1995.1032. [MD]

Fox Tree, J. E. (2001) Listeners' uses of um anduh in speech comprehension. *Memory & Cognition* **29**(2):320–26. doi:10.3758/BF03194926. [MD]

Frank, M. C. & Goodman, N. D. (2012) Predicting pragmatic reasoning in language games. *Science* **336**(6084):998. doi:10.1126/science.1218633. [aFL, MD]

Frank, R. H. (1988) *Passions within reason: The strategic role of the emotions*. WW Norton & Co. [rFL]

Frauchiger, D. & Renner, R. (2018) Quantum theory cannot consistently describe the use of itself. *Nature Communications* **9**:Article 3711. [CM]

Friedman, M. & Savage, L. J. (1948) The utility analysis of choices involving risk. *The Journal of Political Economy* **56**(4):279–304. doi:10.1086/256692. [aFL]

Friedman, M. & Savage, L. J. (1952) The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy* **60**(6):463–74. [aFL]

Frijda, N. H., Kuipers, P. & ter Schure, E. (1989) Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology* **57**(2):212–28. Available at: http://doi.org/10.1037/0022-3514.57.2.212. [EMR]

Friston, K. (2010) The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11**(2):127–38. Available at: https://doi.org/10.1038/nrn2787. [aFL, JET]

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. (2017) Active inference: A process theory. *Neural Computation* **29**(1):1–49. Available at: https://doi.org/10.1162/NECO_a_00912. [JET]

Fudenberg, D., Strack, P. & Strzalecki, T. (2018) *Speed, accuracy, and the optimal timing of choices* (Working paper). MIT Press. [aFL]

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H. & Dingemanse, M. (2017) Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In: *Proceedings of the 39th annual meeting of the cognitive science society*, ed. G. Gunzelmann, A. Howes, T. Tenbrink & E. Davelaar, pp. 2055–60. [MD]

Gabaix, X. (2014) A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics* **129**(4):1661–710. doi:10.1093/qje/qju024. [aFL]

Gabaix, X. (2016) *Behavioral macroeconomics via sparse dynamic programming*. NBER Working Paper No. w21848. National Bureau of Economic Research. [aFL]

Gabaix, X. (2017) *Behavioral inattention*. NBER Working Paper No. 24096. National Bureau of Economic Research. [aFL]

Gabaix, X. & Laibson, D. (2005) *Bounded rationality and directed cognition* (NBER and Harvard working paper). National Bureau of Economic Research. [aFL]

Gabaix, X., Laibson, D., Moloche, G. & Weinberg, S. (2006) Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review* **96**(4):1043–68. doi:10.1257/aer.96.4.1043. [aFL]

Gagne, C., Dayan, P. & Bishop, S. J. (2018) When planning to survive goes wrong: Predicting the future and replaying the past in anxiety and PTSD. *Current Opinion in Behavioral Sciences* **24**:89–95. [rFL]

Gaissmaier, W. & Schooler, L. J. (2008) The smart potential behind probability matching. *Cognition* **109**(3):416–22. [AS]

Galileo, G. (1632/2001) *Dialogue concerning the two Chief World Systems*. Translated by S. Drake. Random House. [CJK]

Ganguli, D. & Simoncelli, E. P. (2014) Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation* **26**(10):2103–34. doi:10.1162/NECO_a_00638. [aFL, WJM]

Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. (2015) Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349** (6245):273–78. doi:10.1126/science.aac6076. [aFL]

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

Gibson, E. J. (1982) The concept of affordances in development: The renascence of functionalism. In: *The concept of development: The Minnesota symposia on child psychology, vol. 15*, ed. W. A. Collins, pp. 55–81. Lawrence Erlbaum. [KPK]

Gigerenzer, G. (2010) *Rationality for mortals: How people cope with uncertainty*, 1st edition. Oxford University Press. [CJK]

Gigerenzer, G. (2015) On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology* **6**:363–83. doi:10.1007/s13164- 015-0248-1. [aFL]

Gigerenzer, G., Fiedler, K. & Olsson, H. (2012) Rethinking cognitive biases as environmental consequences. In: *Ecological rationality: Intelligence in the world*, ed. P. M. Todd, G. Gigerenzer & ABC Research Group, pp. 80–110. Oxford University Press. [aFL]

Gigerenzer, G. & Gaissmaier, W. (2011) Heuristic decision making. *Annual Review of Psychology* **62**(1):451–82. doi:10.1146/annurev-psych-120709-145346. [aFL]

Gigerenzer, G. & Goldstein, D. G. (1996) Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**(4):650–69. doi:10.1037/0033-295X.103.4.650. [aFL]

Gigerenzer, G. & Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* **102**(4):684–704. doi:10.1037/0033-295X.102.4.684. [aFL]

Gigerenzer, G. & Selten, R., eds. (2001) *Bounded rationality: The adaptive toolbox*. MIT Press. [AWS]

Gigerenzer, G. & Selten, R. (2002) *Bounded rationality: The adaptive toolbox*. MIT Press. [aFL]

Gigerenzer, G., Todd, P. M. & ABC Research Group. (1999) *Simple heuristics that make us smart*. Oxford University Press. [aFL]

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J. & Wheatley, T. P. (1998) Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology* **75**:617–38. [EM]

Gilovich, T., Griffin, D. & Kahneman, D., eds. (2002) *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press. [aFL, KPK]

Gintis, H. (2009) *The bounds of reason*. Princeton University Press. [MC]

Giszter, S. F. (2015) Motor primitives – New data and future questions. *Current Opinion Neurobiology* **33**:156–65. [ND]

Gleason, A. M. (1957) Measures on the closed subspaces of a Hilbert space. *Journal of Mathematics and Mechanics* **6**(6):885–93. [HA]

Glymour, C. (1987) Android epistemology and the frame problem. In: *The robot's dilemma: The frame problem in artificial intelligence*, ed. Z. W. Pylyshyn, pp. 63–75. Ablex. [aFL]

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I. & Pine, J. M. (2001) Chunking mechanisms in human learning. *Trends in Cognitive Sciences* **5**(6):236–43. doi:10.1016/S1364-6613(00)01662-4. [aFL]

Goffman, E. (1967) *Interaction ritual*. Aldine. [MD]

Gold, J. I. & Shadlen, M. N. (2007) The neural basis of decision making. *Annual Review of Neuroscience* **30**:535–74. [rFL]

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H. & Dahl, R. E. (2017) Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences* **114**(30):7892–99. [KP]

Gopnik, A., Sobel, D. M., Schulz, L. E. & Glymour, C. (2001) Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology* **37**(5):620–29. [VRB]

Gori, M., Del Viva, M., Sandini, G. & Burr, D. C. (2008) Young children do not integrate visual and haptic form information. *Current Biology* **18**(9):694–98. doi:10.1016/j.cub.2008.04.036. [VRB]

Gottlieb, J., Oudeyer, P.-Y., Lopes, M. & Baranes, A. (2013) Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences* **17**(11):585–93. doi:10.1016/j.tics.2013.09.001. [aFL]

Green, D. W. & Abutalebi, J. (2013) Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology* **25**(5):515–30. [JH]

Griffiths, T., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* **14**(8):357–64. doi:10.1016/j.tics.2010.05.004. [aFL]

Griffiths, T. L., Chater, N., Norris, D. & Pouget, A. (2012) How the Bayesians got their beliefs (and what those beliefs actually are): Comments on Bower and Davis. *Psychological Bulletin* **138**:415–22. [MC]

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008) Bayesian models of cognition. In: *The Cambridge handbook of computational cognitive modeling*, ed. R. Sun. Cambridge University Press. [aFL]

Griffiths, T. L., Lieder, F. & Goodman, N. D. (2015) Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* **7**(2):217–29. doi:10.1111/tops.12142. [aFL]

Griffiths, T. L. & Tenenbaum, J. B. (2001) Randomness and coincidences: Reconciling intuition and probability theory. In: *Proceedings from The 23rd Annual Conference of the Cognitive Science Society* (Edinburgh, Scotland), pp. 370–75. Cognitive Science Society. [aFL]

Griffiths, T. L. & Tenenbaum, J. B. (2006) Optimal predictions in everyday cognition. *Psychological Science* **17**(9):767–73. doi:10.1111/j.1467-9280.2006.01780.x. [aFL]

Griffiths, T. L. & Tenenbaum, J. B. (2009) Theory-based causal induction. *Psychological Review* **116**(4):661–716. doi:10.1037/a0017201. [aFL]

Griffiths, T. L., Vul, E. & Sanborn, A. N. (2012) Bridging levels of analysis for probabilistic models of cognition. *Current Direction in Psychological Science* **21**(4):263–68. doi:10.1177/0963721412447619. [aFL]

Griskevicius, V., Shiota, M. N. & Neufeld, S. L. (2010) Influence of different positive emotions on persuasion processing: A functional evolutionary approach. *Emotion* **10** (2):190–206. [KP]

Guillory, S. A. & Bujarski, K. A. (2014) Exploring emotions using invasive methods: Review of 60 years of human intracranial electrophysiology. *Social Cognitive and Affective Neuroscience* **9**(12):1880–89. Available at: https://doi.org/10.1093/scan/nsu002. [JET]

Gul, S., Krueger, P. M., Callaway, F., Griffiths, T. L. & Lieder, F. (2018) Discovering rational heuristics for risky choice. *KogWis 2018* [Abstract]. [rFL]

Hahn, U. & Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review* **114**(3):704–32. doi:10.1037/0033-295X.114.3.704. [aFL]

Hahn, U. & Warren, P. A. (2009) Perceptions of randomness: Why three heads are better than four. *Psychological Review* **116**(2):454–61. doi:10.1037/a0017522. [aFL]

Hájek, A. (2008) Arguments for – or against – probabilism? *British Journal for the Philosophy of Science* **59**:793–819. [MC]

Hall, B. (2007) *Strickberger's evolution*, 4th edition. Jones and Bartlett Publishers. [CJK]

Halpern, J. Y. & Pass, R. (2015) Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory* **156**(C):246–68. doi:10.1016/j.jet.2014.04.007. [arFL]

Hammond, K. R. (2000) Coherence and correspondence theories in judgment and decision making. In: *Judgment and decision making: An interdisciplinary reader*, 2nd edition, ed. T. Conolly, K. R. Hammond & H. Arkes, pp. 53–65. Cambridge University Press. [AS]

Harman, G. (2013) Rationality. In: International Encyclopedia of Ethics. ed. H. LaFollette, J. Deigh & S. Stroud. Blackwell Publishing Ltd. [aFL]

Harrison, G. & Ross, D. (2017) The empirical adequacy of cumulative prospect theory and its implications for normative assessment. *Journal of Economic Methodology* **24**:150–65. [DRo]

Harrison, G. & Swarthout, J. T. (2016) *Cumulative prospect theory in the laboratory: A reconsideration* (CEAR Working Paper No. 2016-05). Center for Economic Analysis of Risk, Robinson College of Business, Georgia State University. Available at: https://cear.gsu.edu/files/2016/06/WP_2016_05_Cumulative-Prospect-Theory-in-the-Laboratory-A-Reconsideration_MAR-2017.pdf. [DRo]

Hartwell, L. H., Hood, L., Goldberg, M., Reynolds, A. E. & Silver, L. (2011) *Genetics: From genes to genomes. 4th Edition*. McGraw Hill. [CJK]

Haselton, M. G. & Nettle, D. (2006) The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review* **10**(1):47–66. [aFL]

Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. (2017) Neuroscience-inspired artificial intelligence. *Neuron* **95**(2):245–58. doi:10.1016/j.neuron.2017.06.011. [aFL]

Hauser, T. U., Moutoussis, M., Purg, N., Dayan, P. & Dolan, R. J. (2018) Beta-blocker propranolol modulates decision urgency during sequential information gathering. *Journal of Neuroscience* **38**(32):7170–78. Available at: http://doi.org/10.1523/JNEUROSCI.0192-18.2018. [EMR]

Hawkins, J. A. (2004) *Efficiency and complexity in grammars*. Oxford University Press. [aFL]

Hedström, P. & Stern, C. (2008) Rational choice and sociology. In: *The new Palgrave dictionary of economics* (2nd edition), ed. S. N. Durlauf & L. E. Blume. Palgrave Macmillan. [aFL]

Herrnstein, R. J. (1961) Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behaviour* **4**:267–72. doi:10.1901/jeab.1961.4-267. [aFL]

Herron, J. C. & Freeman, S. (2013) *Evolutionary analysis*, 5th edition. Pearson. [CJK]

Hertwig, R. & Hoffrage, U. (2013) *Simple heuristics in a social world*. Oxford University Press. [aFL]

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005) Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **31**(4):621. doi:10.1037/0278-7393.31.4.621. [aFL]

Heyes, C. (2018) *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press. [MD]

Hilbert, M. (2012) Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin* **138**(2):211–37. doi:10.1037/a0025940. [aFL]

Hirsch, J., Adam Noah, J., Zhang, X., Dravida, S. & Ono, Y. (2018) A cross-brain neural mechanism for human-to-human verbal communication. *Social Cognitive and Affective Neuroscience* **13**(9):907–20. doi:10.1093/scan/nsy070. [MD]

Hoffman, B. L., Felter, E. M., Chu, K. H., Shensa, A., Hermann, C., Wolynn, T. & Primack, B. A. (2019) It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. *Vaccine* **37**(16):2216–23. [KPK]

Holmes, P. & Cohen, J. D. (2014) Optimality and some of its discontents: Successes and shortcomings of existing models for binary decisions. *Topics in Cognitive Science* **6** (2):258–78. doi:10.1111/tops.12084. [aFL]

Horvitz, E. J. (1987) Reasoning about beliefs and actions under computational resource constraints. *In:* Proceedings of the third conference on uncertainty in artificial intelligence, pp. 429-44. [aFL]

Horvitz, E. J. (1990) Computation and action under bounded resources. PhD Dissertation, Stanford University. [aFL]

Horvitz, E. J., Cooper, G. F. & Heckerman, D. E. (1989) Reflection and action under scarce resources: Theoretical principles and empirical study. In: *Proceedings from IJCAI-89: The 11th international joint conference on artificial intelligence* (Detroit, Michigan), *Volume* 2, pp. 1121–27. [aFL]

Houston, A. I. & McNamara, J. M. (1999) *Models of adaptive behaviour: An approach based on state*. Cambridge University Press. [aFL]

Howes, A., Duggan, G. B., Kalidindi, K., Tseng, Y. -C. & Lewis, R. L. (2016) Predicting short-term remembering as boundedly optimal strategy choice. *Cognitive Science* **40** (5):1192–223. doi:10.1111/cogs.12271. [aFL]

Howes, A., Lewis, R. L. & Vera, A. H. (2009) Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review* **116**(4):717–51. doi:10.1037/a0017187 [RLL, aFL]

Howes, A., Warren P. A., Farmer, G., El-Deredy, W. & Lewis, R. L. (2016) Why contextual preference reversals maximize expected value. *Psychology Review* **123**(4):368–91. doi:10.1037/a0039996. [aFL]

Hudson Kam, C. L. & Newport, E. L. (2005) Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* **1**(2):151–95. [VRB]

Hume, D. (1739–40/2000) *A treatise of human nature*, ed. D. F. Norton & M. J. Norton. Oxford University Press. [CJK]

Hutchins, E. (1995) *Cognition in the wild*. MIT Press. [MD]

Hutchinson, B. & Barrett, L. F. (2019) The power of predictions: An emerging paradigm for psychological research. *Current Directions in Psychological Science* **28**(3):280–91. Available at: https://doi.org/10.1177/0963721419831992. [JET]

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P. & Roiser, J. P. (2015) Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* **112**(10):3098–103. doi:10.1073/pnas.1414219112. [aFL]

Huys, Q. J. M. & Renz, D. (2017) A formal valuation framework for emotions and their control. *Biological Psychiatry* **82**(6):413–20. Available at: http://doi.org/10.1016/j.biopsych.2017.07.003. [EMR]

Icard, T. (2014) Toward boundedly rational analysis. In: *Proceedings from the 36th annual conference of the Cognitive Science Society* (Quebec, Canada), Volume 1, pp. 637–42. Cognitive Science Society. [aFL]

Icard, T. & Goodman, N. D. (2015) A resource-rational approach to the causal frame problem. In: *Proceedings from the 37th annual meeting of the Cognitive Science Society* (Pasadena, CA). Cognitive Science Society. [aFL]

Jain, Y. R., Gupta, S., Rakesh, V., Dayan, P., Callaway, F. & Lieder, F. (in press) Testing models of how people learn how to plan. [rFL]

Jazayeri, M. & Shadlen, M. N. (2010) Temporal context calibrates interval timing. *Nature Neuroscience* **13**(8):1020–26. doi:10.1038/nn.2590. [VRB]

Jern, A., Chang, K.-M. K. & Kemp, C. (2014) Belief polarization is not always irrational. *Psychological Review* **121**(2):206–24. [EM, rFL]

Johnstone, R. A., Dall, S. R. X. & Dukas, R. (2002) Behavioural and ecological consequences of limited attention. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **357**. Available at: http://doi.org/10.1098/rstb.2002.1063. [aFL]

Jusczyk, P. W. & Aslin, R. N. (1995) Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* **29**(1):1–23. doi:10.1006/cogp.1995.1010. [VRB]

Juslin, P., Nilsson, H. & Winman, A. (2009) Probability theory, not the very guide of life. *Psychological Review* **116**(4):856–74. [ANS]

Kahneman, D. (2003) Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review* **93**(5):1449-75. doi:10.1257/000282803322655392. [aFL]

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman. Cambridge University Press. doi:10.1017/CBO9780511808098.004. [aFL]

Kahneman, D. & Frederick, S. (2005) A model of heuristic judgment. In: *The Cambridge handbook of thinking and reasoning*, ed. K. J. Holyoak & R. G. Morrison, pp. 267–93. Cambridge University Press. [aFL]

Kahneman, D. & Tversky, A. (1972) Subjective probability: A judgment of representativeness. *Cognitive Psychology* **3**(3):430–54. doi:10.1016/0010-0285(72)90016-3. [aFL]

Kahneman, D. & Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2):263–91. doi:10.2307/1914185. [arFL]

Kauffman, S. A. (1993) *The origins of order: Self-organization and selection in evolution*. Oxford University Press. [KPK]

Kemp, C. & Regier, T. (2012) Kinship categories across languages reflect general communicative principles. *Science* **336**(6084):1049–54. doi:10.1126/science.1218811. [aFL]

Kemp, C. & Tenenbaum, J. B. (2008) The discovery of structural form. *Proceedings of the National Academy of Sciences* **105**(31):10687–92. [KP]

Kempson, R., Cann, R., Gregoromichelaki, E. & Chatzikyriakidis, S. (2016) Language as mechanisms for interaction. *Theoretical Linguistics* **42**(3–4):203–76. doi:10.1515/tl-2016-0011. [MD]

Keramati, M., Dezfouli, A. & Piray, P. (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *The Public Library of Science Computational Biology* **7**(5):e1002055, 1–21. doi:10.1371/journal.pcbi.1002055. [aFL]

Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences* **113**(45):12868–73. doi:10.1073/pnas.1609094113. [aFL]

Kersten, A. W. & Earles, J. L. (2001) Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language* **44**(2):250–73. [VRB]

Khaw, M. W., Li, Z. & Woodford, M. (2017) *Risk aversion as a perceptual bias*. NBER Working Paper No. 23294. National Bureau of Economic Research. [aFL]

Khrennikov, A., Basieva, I., Pothos, E. M. & Yamato, I. (2018) Quantum probability in decision making from quantum information representation of neuronal states. *Scientific Reports* **8**(1):16225. [HA]

Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P. (2002) Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition* **83**(2):B35–B42. [VRB]

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., Quigley, K. S., Dickerson, B. C. & Barrett, L. F. (2017) Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behaviour* **1** (5):0069. Available at: https://doi.org/10.1038/s41562-017-0069. [JET]

Klein, C. (2018) Mechanisms, resources, and background conditions. *Biology and Philosophy* **33**:36. [JH]

Knill, D. C. & Pouget, A. (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27**(12):712–19. doi:10.1016/j.tins.2004.10.007. [aFL]

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

Konvalinka, I. & Roepstorff, A. (2012) The two-brain approach: How can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience* **6**:215. doi:10.3389/fnhum.2012.00215. [MD]

Kool, W. & Botvinick, M. M. (2013) The intrinsic cost of cognitive control. *The Behavioral and Brain Sciences* **36**(6):697–98. doi:10.1017/S0140525X1300109X. [aFL]

Körding, K. P. & Wolpert, D. M. (2004) Bayesian integration in sensorimotor learning. *Nature* **427**(6971):244–47. doi:10.1038/nature02169. [aFL, VRB]

Krajbich, I., Lu, D., Camerer, C. & Rangel, A. (2012) The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology* **3**:193. [rFL]

Kreiner, J. (2014) *The social life of hagiography in the Merovingian kingdom*. Cambridge University Press. [HMC]

Krueger, P. M. & Griffiths, T. (2018) Shaping model-free habits with model-based goals. In: *CogSci 2018*. [rFL]

Krueger, P. M., Lieder, F. & Griffiths, T. (2017) Enhancing metacognitive reinforcement learning using reward structures and feedback. In: *CogSci 2017*. [rFL]

Kuhl, P. K. & Meltzoff, A. N. (1982) The bimodal perception of speech in infancy. *Science* **218**(4577):1138–41. [VRB]

Kurkul, K. E. & Corriveau, K. H. (2018) Question, explanation, follow-up: A mechanism for learning from others? *Child Development* **89**(1):280–94. [KP]

Kwon, O.-S. & Knill, D. C. (2013) The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning. *Proceedings of the National Academy of Sciences* **110**(11):E1064–73. doi:10.1073/pnas.1214869110. [VRB]

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2015) Human-level concept learning through probabilistic program induction. *Science* **350**:1332–38. [HA]

Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. (2017) Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**(253):1-72. doi:10.1017/S0140525X16001837. [aFL]

Langley, P., Laird, J. E. & Rogers, S. (2009) Cognitive architectures: Research issues and challenges. *Cognitive Systems Research* **10**(2):141–60. doi:10.1016/j.cogsys.2006.07.004. [aFL]

Larrick, R. P. (2004) Debiasing. In: *Blackwell handbook of judgment and decision making*, ed. D. J. Koehler & N. Harvey, pp. 316–38. Blackwell Publishing. [aFL]

Latty, T. & Beekman, M. (2010) Irrational decision-making in an amoeboid organism: Transitivity and context-dependent preferences. *Proceedings of the Royal Society B: Biological Sciences* **278**(1703): 307–12. [aFL]

Laughlin, S. (1981) A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung C* **36**(9–10):910–12. [WJM]

Lennie, P. (2003) The cost of cortical computation. *Current Biology* **13**(6):493–97. doi:10.1016/S0960-9822(03)00135-0. [aFL]

Lerch, R. A. & Sims, C. R. (2019) Rate-distortion theory and computationally rational reinforcement learning. In: *Proceedings of Reinforcement Learning and Decision Making (RLDM) 2019*, July 7–10, Montreal, Canada. [CJB]

Lestrade, S. (2017) Unzipping Zipf's law. *PLoS One* **12**(8):e0181987. doi:10.1371/journal.pone.0181987. [MD]

Levelt, W. J. M. (1989) *Speaking: From intention to articulation*. MIT Press. [MD]

Levinson, S. C. (2000) *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press. [MD]

Levinson, S. C. (2016) Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences* **20**(1):6–14. doi:10.1016/j.tics.2015.10.010. [MD]

Levy, W. B. & Baxter, R. A. (1996) Energy efficient neural codes. *Neural Computation* **8** (3):531–43. doi:10.1162/neco.1996.8.3.531. [aFL]

Levy, W. B. & Baxter, R. A. (2002) Energy-efficient neuronal computation via quantal synaptic failures. *Journal of Neuroscience* **22**(11):4746–55. [aFL]

Lewis, M. (2008) The emergence of human emotions. In: *Handbook of emotions*, 3rd edition, ed. M. Lewis, J. M. Haviland-Jones & L. F. Barrett, pp. 304–19. Guilford Press. [KP]

Lewis, R. L., Howes, A. & Singh, S. (2014) Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science* **6** (2):279–311. doi:10.1111/tops.12086. [aFL, RLL]

Li, V., Castañon, S. H., Solomon, J. A., Vandormael, H. & Summerfield, C. (2017) Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology* **13**(8):e1005723. [WJM]

Liberman, A. & Chaiken, S. (1992) Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin* **18**:669–79. [EM]

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M. & Combs, B. (1978) Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**(6):551–78. [aFL]

Lieder, F. (2018) Beyond bounded rationality: Reverse-engineering and enhancing human intelligence (Doctoral dissertation). University of California, Berkeley. [aFL, WJM]

Lieder, F., Callaway, F., Krueger, P. M., Das, P., Griffiths, T. L. & Gul, S. (2018a) Discovering and teaching optimal planning strategies. In: *The 14th biannual conference of the German Society for Cognitive Science, GK*. [aFL]

Lieder, F., Griffiths T. L. & Hsu, M. (2018b) Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review* **125**(1):1–32. doi:10.1037/rev0000074.

Lieder, F., Callaway F., Jain, Y. R., Krueger P. M., Das, P., Gul S. & Griffiths, T. L. (2019a) A cognitive tutor for helping people overcome present bias. *RLDM* 2019. doi:10.13140/RG.2.2.10467.20006. [aRF]

Lieder, F., Chen O. X., Krueger, P. M. & Griffiths, T. L. (2019b) Cognitive prostheses for goal achievement. *Nature Human Behavior* **3**:1096–1106. [aFL]

Lieder, F. & Griffiths, T. L. (2017) Strategy selection as rational metareasoning. *Psychological Review* **124**(6):762–94. doi:10.1037/rev0000075. [aFL, AS]

Lieder, F., Griffiths, T. L. & Goodman, N. D. (2012) Burn-in, bias, and the rationality of anchoring. In: *Advances in Neural Information Processing Systems*, vol. **26**, ed. P. Bartlett, F. C. N. Pereira, L. Bottou, C. J. C. Burges & K. Q. Weinberger, pp. 2690–798. Curran Associates, Inc. [aFL]

Lieder, F., Griffiths, T. L. & Hsu, M. (2018b) Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review* **125**(1):1–32. doi:10.1037/rev0000074. [aFL, EMR]

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018c) Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review* **25** (2):775-84. doi:10.3758/s13423-017-1288-6. [aFL]

Lieder, F., Griffiths, T. L., Huys, Q. J. M. & Goodman, N. D. (2018d) The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review* **25**(1):322–49. doi:10.3758/s13423-017-1286-8. [aFL, WJM]

Lieder, F., Hsu, M. & Griffiths, T. L. (2014) The high availability of extreme events serves resource-rational decision-making. in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Cognitive Science Society. [aFL, WJM]

Lieder, F., Krueger, P. M. & Griffiths, T. L. (2017) An automatic method for discovering rational heuristics for risky choice. In: *Proceedings from the 39th annual conference of the Cognitive Science Society* (London, UK), pp. 2567–72. Cognitive Science Society. [aFL]

Lieder, F., Shenhav, A., Musslick, S. & Griffiths, T. L. (2018e) Rational metareasoning and the plasticity of cognitive control. *The Public Library of Science Computational Biology* **14**(4):e1006043. https://doi.org/10.1371/journal.pcbi.1006043. [aFL]

Locke, E. & Latham, G. (2002) Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist* **57**(9):705–17. doi:10.1037/0003-066x.57.9.705. [aFL]

Lohmann, S. (2008) Rational choice and political science. In: *The new Palgrave dictionary of economics*, 2nd edition, ed. S. N. Durlauf & L. E. Blume. Palgrave Macmillan. doi:10.1007/978-1-349-58802-2_1383. [aFL]

Luciana, M. & Nelson, C. A. (1998) The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. *Neuropsychologia* **36** (3):273–93. [VRB]

Machina, M. (2009) Risk, ambiguity, and the rank-dependence axioms. *Journal of Economic Review* **99**(1):385–92. [CM]

MacKenzie, D. (2006) *An engine, not a camera: How financial models shape markets*. MIT Press. [HMC]

Mackowiak, B. & Wiederholt, M. (2009) Optimal sticky prices under rational inattention. *American Economic Review* **99**(3):769–803. [WJM]

Mahowald, K., Fedorenko, E., Piantadosi, S. T. & Gibson, E. (2013) Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* **126**(2):313–18. doi:10.1016/j.cognition.2012.09.010. [aFL]

Mandelbaum, E. (2019) Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language* **34**:141–57. [EM]

Marblestone, A. H., Wayne, G. & Kording, K. P. (2016) Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience* **10**:94. doi:10.3389/fncom.2016.00094. [MC]

Marcus, G. (2008) *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt. [ESD, aFL]

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

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press. [arFL, AS]

Matějka, F. & McKay, A. (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* **105**(1):272–98. doi:10.1257/aer.20130047. [aFL]

Mathews, A. & MacLeod, C. (2005) Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology* **1**(1):167–95. Available at: http://doi.org/10.1146/annurev.clinpsy.1.102803.143916. [EMR]

Maylor, E. A., Chater, N. & Jones, G. V. (2001) Searching for two things at once: Evidence of exclusivity in semantic and autobiographical memory retrieval. *Memory & Cognition* **29**(8):1185–95. [ANS]

McNamara, J. M. & Weissing, F. J. (2010) Evolutionary game theory. In: *Social behaviour: genes, ecology and evolution*, ed. T. Székely, A. J. Moore & J. Komdeur, pp. 88–106. Cambridge University Press. [aFL]

Meyer, D. E. & Kieras, D. E. (1997a) A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological Review* **104**(1):3–65. doi:10.1037/0033-295X.104.1.3. [aFL, CD]

Meyer, D. E. & Kieras, D. E. (1997b) A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review* **104**(4):749–91. doi:10.1037//0033-295X.104.4.749. [aFL, CD]

Mill, J. S. (1882) *A system of logic, ratiocinative and inductive*, 8th edition. Harper and Brothers. [aFL]

Milli, S., Lieder, F. & Griffiths, T. L. (2017) When does bounded-optimal metareasoning favor few cognitive systems? In: *Proceedings from AAAI-17: The 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, vol. 31, 4422–28. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press. [aFL]

Milli, S., Lieder, F. & Griffiths, T. L. (2019) *A rational reinterpretation of dual-process theories*. Preprint. doi:10.13140/RG.2.2.14956.46722/1. [aFL]

Mineka, S., Davidson, M., Cook, M. & Keir, R. (1984) Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology* **93**(4):355–72. Available at: https://doi.org/10.1037/0021-843X.93.4.355. [JET]

Młynarski, W. F. & Hermundstad, A. M. (2018) Adaptive coding for dynamic sensory inference. *Elife* **7**:e32055. [WJM]

Moore, D. A. & Healy, P. J. (2008) The trouble with overconfidence. *Psychological Review* **115**(2):502–17. [aFL]

Moore, G. E. (1903) *Principia ethica*. Cambridge University Press. [CJK]

Moreira, C., Haven, E., Sozzo, S. & Wichert, A. (2018) Process mining with real world financial loan applications: Improving inference on incomplete event logs. *PLoS One* **13**(12):e0207806. [CM]

Mossio, M. & Moreno, A. (2015) Biological autonomy: A philosophical and theoretical enquiry. History, philosophy, and theory of life sciences series. Springer. [KPK]

Musslick, S., Dey, B., Ozcimder, K., Patwary, M. M. A., Willke, T. L. & Cohen, J. D. (2016) Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In: *Proceedings from The 38th Annual Conference of the Cognitive Science Society* (Philadelphia, PA), pp. 1547–52. Cognitive Science Society. [arFL]

Musslick, S., Saxe, A. M., Ozcimder, K., Dey, B., Henselman, G. & Cohen, J. D. (2017) Multitasking capability versus learning efficiency in neural network architectures. In: *Proceedings from The 39th Cognitive Science Society Conference* (London, UK), pp. 829–34. Cognitive Science Society. [arFL]

Nagel, T. (2012) *Mind and cosmos: Why the materialist Neo-Darwinian conception of nature is almost certainly false*, 1st edition. Oxford University Press. [CJK]

Nardini, M., Bedford, R. & Mareschal, D. (2010) Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences* **107**(39):17041–46. doi:10.1073/pnas.1001699107. [VRB]

Nardini, M., Jones, P., Bedford, R. & Braddick, O. (2008) Development of cue integration in human navigation. *Current Biology* **18**(9):689–93. doi:10.1016/j.cub.2008.04.021. [VRB]

Navon, D. & Gopher, D. (1979) On the economy of the human-processing system. *Psychological Review* **86**(3):214–55. doi:10.1037/0033-295X.86.3.214. [aFL]

Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J. & Shimojo, S. (2006) Development of multisensory spatial integration and perception in humans. *Developmental Science* **9**(5):454–64. [VRB]

Netzer, N. (2009) Evolution of time preferences and attitudes toward risk. *American Economic Review* **99**(3):937–55. [WJM]

Neuman, R., Rafferty, A. & Griffiths, T. (2014) A bounded rationality account of wishful thinking. In: *Proceedings from the 36th annual meeting of the Cognitive Science Society*. Cognitive Science Society. [aFL]

Newell, A. (1990) *Unified theories of cognition*. Harvard University Press. [CD, RLL]

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press

Newell, A., Shaw, J. C. & Simon, H. A. (1958) Elements of a theory of human problem solving. *Psychological Review* **65**(3):151–66. doi:10.1037/h0048495. [aFL]

Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [aFL]

Newell, B. R. (2005) Re-visions of rationality? *Trends in Cognitive Sciences* **9**(1):11–15. [AS]

Newell, B. R. & Shanks, D. R. (2014) Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences* **37**(1):1–19. [AS]

Newport, E. L. (1990) Maturational constraints on language learning. *Cognitive Science* **14** (1):11–28. doi:10.1016/0364-0213(90)90024-Q. [VRB]

Niven, J. E. & Laughlin, S. B. (2008) Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology* **211**(11):1792–804. Available at: https://doi.org/10.1242/jeb.017574. [aFL, JET]

Nobandegani, A. (2017) *The minimalist mind: On minimality in learning, reasoning*. McGill-Queen's University Press. [aFL]

Nobandegani, A. S., Castanheira, K. da S., Otto, A. R. & Shultz, T. R. (2018) Over-representation of extreme events in decision-making: A rational metacognitive account. In: *Proceedings from the 40th annual conference of the Cognitive Science Society*, pp. 2394–99. Cognitive Science Society. [aFL]

Nobandegani, A. S. & Psaromiligkos, I. N. (2017) The causal frame problem: An algorithmic perspective. In: *Proceedings from the 39th annual conference of the Cognitive Science Society* (London, UK), pp. 2567–72. Cognitive Science Society. [aFL]

Oaksford, M. & Chater, N. (1994) A rational analysis of the selection task as optimal data selection. *Psychological Review* **101**(4):608–31. doi:10.1037/0033-295X.101.4.608. [aFL, CD, ESD]

Oaksford, M. & Chater, N. (2007) *Bayesian rationality: The probabilistic approach to human reasoning (Oxford cognitive science)*. Oxford University Press. [aFL]

Okasha, S. (2013) The evolution of Bayesian updating. *Philosophy of Science* **80**(5):745–57. [DS]

Olshausen, B. A. & Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**(6583):607–09. doi:10.1038/381607a0. [aFL, WJM]

Olshausen, B. A. & Field, D. J. (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* **37**(23):3311–25. doi:10.1016/S0042-6989(97)00169-7. [aFL]

Olshausen, B. A. & Field, D. J. (2004) Sparse coding of sensory inputs. *Current Opinion in Neurobiology* **14**(4):481–87. doi:10.1016/j.conb.2004.07.007. [aFL]

Orhan, A. E., Sims, C. R., Jacobs, R. A. & Knill, D. C. (2014) The adaptive nature of visual working memory. *Current Directions in Psychological Science* **23**(3):164–70. doi:10.1177/0963721414529144. [aFL]

Park, I. M. & Pillow, J. W. (2017) Bayesian efficient coding. bioRxiv. Preprint. doi:10.1101/178418. [WJM]

Pashler, H. E. & Sutherland, S. (1998) *The psychology of attention*, vol. **15**. MIT Press. [aFL]

Pavlov, I. P. (1927) *Conditioned reflexes*. Oxford University Press. [KPK]

Payne, J. W., Bettman, J. R. & Johnson, E. J. (1993) *The adaptive decision maker*. Cambridge University Press. [aFL]

Peil, K. T. (2012) Emotion: A self-regulatory sense? EFS International. Available at: http://www.academia.edu/7208004/Emotion_. [KPK]

Peil, K. T. (2014) The self-regulatory sense. *Global Advances in Health Medicine* **3**(2):80–108. [KPK]

Peil, K. T. (2017) The resonant biology of emotion. *Constructivist Foundations* **12**(2):232–33. [KPK]

Pert, C. (1998) *The molecules of emotion*. Touchstone. [KPK]

Peters, M. A. K., Ma, W. J. & Shams, L. (2016) The size-weight illusion is not anti-Bayesian after all: A unifying Bayesian account. *Peer J* **4**:e2124. [EM]

Petrini, K., Remark, A., Smith, L. & Nardini, M. (2014) When vision is not an option: Children's integration of auditory and haptic information is suboptimal. *Developmental Science* **17**(3):376–87. doi:10.1111/desc.12127. [VRB]

Pettit, M. (2017) The great cat mutilation: Sex, social movements and the utilitarian calculus in 1970s New York City. *BJHS: Themes* **2**:57–78. [HMC]

Piantadosi, S. T., Tily, H. & Gibson, E. (2011) Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* **108** (9):3526–29. doi:10.1073/pnas.1012551108. [aFL]

Piantadosi, S. T., Tily, H. & Gibson, E. (2012) The communicative function of ambiguity in language. *Cognition* **122**(3):280–91. doi:10.1016/j.cognition.2011.10.004. [MD]

Piccinini, G. & Schulz, A. (2019) The ways of altruism. *Evolutionary Psychological Science* **5**:58–70. [AWS]

Pittendrigh, C. (1958) Adaptation, natural selection, and behavior. In: *Behavior and evolution*, ed. A. Roe & G. G. Simpson. Yale University Press. [CJK]

Plous, S. (1991) Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology* **21**:1058–82. [EM]

Plous, S. (1993) *The psychology of judgment and decision making*. McGraw-Hill. [ESD]

Polania, R., Woodford, M. & Ruff, C. C. (2019) Efficient coding of subjective value. *Nature Neuroscience* **22**(1):134. [aFL]

Pontzer, H. (2015) Energy expenditure in humans and other primates: A new synthesis. *Annual Review of Anthropology* **44**(1):169–87. Available at: https://doi.org/10.1146/annurev-anthro-102214-013925. [JET]

Pothos, E. M. & Busemeyer, J. R. (2009) A quantum probability model explanation for violations of "rational" decision theory. *Proceedings of the Royal Society B: Biological Sciences* **276**(1665):2171–78. [CM]

Pothos, E. M. & Busemeyer, J. R. (2013) Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences* **36**(3):255–74. [CM, HA]

Radley, J. J., Kabbaj, M., Jacobson, L., Heydendael, W., Yehuda, R. & Herman, J. P. (2011) Stress risk factors and stress-related pathology: Neuroplasticity, epigenetics and endophenotypes. Stress **14**(5):481–97. [KPK]

Rahnev, D. & Denison, R. N. (2018a) Suboptimality in perceptual decision making. *Behavioral and Brain Sciences* **41**:e223, 1–66. doi:10.1017/S0140525X18000936. [DRa, MC]

Rahnev, D. & Denison, R. N. (2018b) Behavior is sensible but not globally optimal: Seeking common ground in the optimality debate. *Behavioral and Brain Sciences* **41**:e251. Available at: https://www.cambridge.org/core/product/identifier/S0140525X18002121/type/journal_article [Accessed January 10, 2019]. [DRa]

Ramsay, F. (1931) *The foundations of mathematics and other logical essays*. Harcourt Brace and Company. [DS]

Rao, R. P. N. & Ballard, D. H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**(1):79–87. Available at: https://doi.org/10.1038/4580. [JET]

Ratcliff, R. (1978) A theory of memory retrieval. *Psychological Review* **85**(2):59–108. doi:10.1037/0033-295X.85.2.59. [aFL]

Ratcliff, R. & McKoon, G. (2008) The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation* **20**(4):873–922. [ANS]

Raymer, E. J. (2013) A man of his time: Thorstein Veblen and the University of Chicago Darwinists. *Journal of the History of Biology* **46**(4):669–98. [HMC]

Regier, T., Kay, P. & Khetarpal, N. (2007) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* **104**(4):1436–41. doi:10.1073/pnas.0610341104. [aFL]

Reis, R. (2006) Inattentive consumers. *Journal of Monetary Economics* **53**(8):1761–1800. doi:10.3386/w10883. [aFL]

Risko, E. F. & Gilbert, S. J. (2016) Cognitive offloading. *Trends in Cognitive Sciences* **20** (9):676–88. doi:10.1016/j.tics.2016.07.002. [MD]

Roberts, S. G. & Levinson, S. C. (2017) Conversation, cognition and cultural evolution: A model of the cultural evolution of word order through pressures imposed from turn taking in conversation. *Interaction Studies* **18**(3):404–31. doi:10.1075/is.18.3.06rob [MD]

Robinson, D. N. (2007) *Consciousness and mental life*, 1st edition. Columbia University Press. [CJK]

Robson, A. & Whitehead, L. A. (2016) *Rapidly adaptive hedonic utility*. Working paper. Simon Fraser University. [WJM]

Robson, A. J. (2001) The biological basis of economic behavior. *Journal of Economic Literature* **39**(1):11–33. [WJM]

Rolls, E. T. (2013) *Emotion and decision-making explained*. Oxford University Press. [DS]

Rowlands, M. (2010) *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press. [KPK]

Rozenblit, L. & Keil, F. (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* **26**(5):521–62. doi:10.1207/s15516709cog2605_1. [aFL]

Rumelhart, D. E. & McClelland, J. L. (1987) Parallel distributed processing, *vol. 1*. MIT Press. [aFL]

Ruse, M. (2017) *On purpose*. Princeton University Press. [CJK]

Russell, S. J. (1997) Rationality and intelligence. *Artificial Intelligence* **94**(1–2):57–77. doi:10.1016/S0004-3702(97)00026-X. [aFL]

Russell, S. J. & Subramanian, D. (1995) Provably bounded-optimal agents. *Journal of Artificial Intelligence Research* **2**(1):575–609. doi: 10.1613/jair.133. [aFL, RLL]

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996) Statistical learning by 8-month-old infants. *Science* **274**(5294):1926–28. doi:10.1126/science.274.5294.1926. [VRB]

Salvucci, D. D. & Taatgen, N. A. (2010) *The multitasking mind*. Oxford University Press. [CD]

Samuels, R., Stich, S. & Bishop, M. (2002) Ending the rationality wars: How to make disputes about human rationality disappear. In: *Common sense, reasoning and rationality*, ed. R. Elio, pp. 236–68. Oxford University Press. [MC]

Sanborn, A. N. & Chater, N. (2016) Bayesian brains without probabilities. *Trends in Cognitive Sciences* **20**(12):883–93. doi:10.1016/j.tics.2016.10.003. [aFL, ANS]

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010) Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* **117** (4):1144–67. doi:10.1037/a0020511. [aFL, HA]

Sanjurjo, A. (2017) Search with multiple attributes: Theory and empirics. *Games and Economic Behavior* **104**:535–62. doi:10.2139/ssrn.2460129. [aFL]

Savage, L. J. (1954) *The foundations of statistics*. Wiley. [DS]

Scherer, K. R. (2009) The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion* **23**(7):1307–51. Available at: http://doi.org/10.1080/02699930902928969. [EMR]

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T. & Vogeley, K. (2013) Toward a second-person neuroscience. *Behavioral and Brain Sciences* **36** (4):393–414. doi:10.1017/S0140525X12000660. [MD]

Schneider, D. W. & Anderson, J. R. (2012) Modeling fan effects on the time course of associative recognition. *Cognitive Psychology* **64**:127–60. [CD]

Scholz, J. P. & Schoner, G. (1999) The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research* **126**:289–306. [ND]

Schulz, A. (2018) *Efficient cognition: The evolution of representational decision making*. MIT Press. [AWS]

Schulze, C. & Newell, B. R. (2016) More heads choose better than one: Group decision making can eliminate probability matching. *Psychonomic Bulletin & Review* **23**:907–14. [AS]

Schwarz, N. (2002) Situated cognition and the wisdom of feelings: Cognitive tuning. In: *The wisdom in feelings*, ed. L. Feldman Barrett & P. Salovey, pp. 144–66. Guilford Press. [KP]

Schwarz, N. & Clore, G. L. (2007) Feelings and phenomenal experiences. In: *Social psychology: A handbook of basic principles*, 2nd edition, ed. E. T. Higgins & A. Kruglanski, pp. 433–65. Guilford Press. [KP]

Sedlmeier, P. & Gigerenzer, G. (2001) Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* **130**(3):380–400. doi:10.1037//0096-3445.130.3.380. [aFL]

Segev, Y., Musslick, S., Niv, Y. & Cohen, J. D. (2018) Efficiency of learning vs. processing: Towards a normative theory of multitasking. In: *Proceedings from the 40th annual conference of the Cognitive Science Society* (Madison, WI). Cognitive Science Society. [arFL]

Selye, H. (1957/1978) *The stress of life*. McGraw-Hill. [KPK]

Sengupta, B., Stemmler, M. B. & Friston, K. J. (2013) Information and efficiency in the nervous system – a synthesis. *PLoS Computational Biology* **9**(7):e1003157. Available at: https://doi.org/10.1371/journal.pcbi.1003157. [JET]

Seth, A. K. (2015) The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In: *Open mind*, ed. T. Metzinger & J. M. Windt, pp. 1–24. MIND Group. Available at: http://www.open-mind.net/DOI?isbn=9783958570108. [JET]

Shadlen, M. N. & Shohamy, D. (2016) Decision making and sequential sampling from memory. *Neuron* **90**(5):927–39. [ANS, rFL]

Shadmehr, R., Smith, M. A. & Krakauer, J. W. (2010) Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience* **33**(1):89–108. Available at: https://doi.org/10.1146/annurev-neuro-060909-153135. [JET]

Shafir, S., Waite, T. A. & Smith, B. H. (2002) Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). *Behavioral Ecology and Sociobiology* **51**(2):180–87. [aFL]

Shanks, D., Tunney, R. & McCarthy, J. (2002) A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making* **15**(3):233–50. doi:10.1002/bdm.413. [aFL]

Shannon, C. & Weaver, W. (1949/1964) *The mathematical theory of communication*, 10th edition. The University of Illinois Press. [JET]

Shaw, M. L. & Shaw, P. (1977) Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance* **3**(2):201. [WJM]

Shenhav, A., Botvinick, M. M. & Cohen, J. (2013) The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* **79**(2):217–40. doi:10.1016/j.neuron.2013.07.007. [aFL]

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D. & Botvinick, M. M. (2017) Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience* **40**:99–124. doi:10.1146/annurev-neuro-072116-031526. [aFL, RLL]

Shimansky, Y. P. & Rand, M. K. (2013) Two-phase strategy of controlling motor coordination determined by task performance optimality. *Biological Cybernetics* **107**:107–29. [ND]

Shrager, J. & Siegler, R. S. (1998) SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science* **9**(5):405–10. [aFL]

Shugan, S. M. (1980) The cost of thinking. *Journal of Consumer Research* **7**(2):99–111. doi:10.1086/208799. [aFL]

Siegler, R. & Jenkins, E. A. (1989) *How children discover new strategies*. Psychology Press. [aFL]

Simon, H. A. (1955) A behavioral model of rational choice. *The Quarterly Journal of Economics* **69**(1):99–118. doi:10.2307/1884852. [aFL]

Simon, H. A. (1956) Rational choice and the structure of the environment. *Psychological Review* **63**(2):129–38. doi:10.1037/h0042769. [aFL]

Simon, H. A. (1982) Models of bounded rationality: Empirically grounded economic reason, *vol. 3*. MIT Press. [aFL]

Simon, H. A. (1996) *The sciences of the artificial*. MIT Press. [JH]

Sims, C. A. (2003) Implications of rational inattention. *Journal of Monetary Economics* **50**(3):665–90. doi:10.1016/S0304-3932(03)00029-1. [aFL, WJM]

Sims, C. A. (2006) Rational inattention: Beyond the linear-quadratic case. *American Economic Review* **96**(2):158–63. doi:10.1257/000282806777212431. [aFL]

Sims, C. R. (2016) Rate-distortion theory and human perception. *Cognition* **152**:181–98. doi:10.1016/j.cognition.2016.03.020. [aFL, CJB]

Sims, C. R. (2018) Efficient coding explains the universal law of generalization in human perception. *Science* **360**:6389, 652–56. [CJB]

Sims, C. R., Jacobs, R. A. & Knill, D. C. (2012) An ideal observer analysis of visual working memory. *Psychological Review* **119**(4):807–30. doi:10.1037/a0029856. [aFL, CJB, WJM]

Sims, C. R., Neth, H., Jacobs, R. A. & Gray, W. D. (2013) Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review* **120**:139–54. [CD]

Singh, S., Lewis, R. L., Barto, A. G. & Sorg, J. (2010) Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development* **2**(2):70–82. [RLL]

Slovic, P., Finucane, M., Peters, E. & MacGregor, D. G. (2002) The affect heuristic. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 397–420. Cambridge University Press. [KPK]

Smith, V. L. (2008) *Rationality in economics: Constructivist and ecological forms*. Cambridge University Press. [MC]

Smith, V. L. (2009) *Rationality in economics*. Cambridge Books. [WJM]

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y. & Botvinick, M. M. (2014) Optimal behavioral hierarchy. *The Public Library of Science Computational Biology* **10**(8):e1003779. doi:10.1371/journal.pcbi.1003779. [aFL]

Sorg, J., Singh, S. & Lewis, R. L. (2010) Internal rewards mitigate agent boundedness. In: *Proceedings of the International Conference on Machine Learning (ICML)*, Haifa, Israel. [RLL]

Sosis, C. & Bishop, M. (2014) Rationality. *Wiley Interdisciplinary Reviews: Cognitive Science* **5**(1):27–37. doi:10.1002/wcs.1263. [aFL]

Spicer, J. & Sanborn, A. N. (2019) What does the mind learn? A comparison of human and machine learning representations. *Current Opinion in Neurobiology* **55**:97–102. [ANS]

Spurrett, D. (2019) The descent of preferences. *British Journal for the Philosophy of Science* axz020. Available at: https://doi.org/10.1093/bjps/axz020. [DS]

Stanovich, K. E. (2011) *Rationality and the reflective mind*. Oxford University Press. [aFL]

Steiner, J. & Stewart, C. (2016) *Perceiving prospects properly*. American Economic Review *106(7):1601–31*. [WJM]

Sterelny, K. (2003) *Thought in a hostile world*. Blackwell. [DRo, DS]

Sterelny, K. (2012) *The evolved apprentice*. MIT Press. [DRo]

Sterling, P. (2004) Principles of allostasis: Optimal design, predictive regulation, pathophysiology, and rational therapeutics. In: *Allostasis, homeostasis, and the costs of physiological adaptation*, ed. J. Schulkin, pp. 17–64. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9781316257081.004. [JET]

Sterling, P. (2012) Allostasis: A model of predictive regulation. *Physiology & Behavior* **106**(1):5–15. Available at: https://doi.org/10.1016/j.physbeh.2011.06.004. [JET]

Sterling, P. & Eyer, J. (1988) Allostasis: A new paradigm to explain arousal pathology. In: *Handbook of life stress, cognition and health*, ed. S. Fisher & J. Reason, pp. 629–49. John Wiley & Sons. [JET]

Sterling, P. & Laughlin, S. (2015) *Principles of neural design*. MIT Press. [aFL, JET]

Sternberg, S. (1966) High-speed scanning in human memory. *Science* **153**(3736):652–54. doi:10.1126/science.153.3736.652. [aFL]

Stewart, N. (2009) Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology* **62**(6):1041–62. doi:10.1080/17470210902747112. [aFL]

Stewart, N., Chater, N. & Brown, G. D. A. (2006) Decision by sampling. *Cognitive Psychology* **53**(1):1–26. doi:10.1016/j.cogpsych.2005.10.003. [aFL]

Stigler, G. J. (1961) The economics of information. *Journal of Political Economy* **69**(3):213–25. [aFL, WJM]

Stocker, A., Simoncelli, E. & Hughes, H. (2006) Sensory adaptation within a Bayesian framework for perception. In: *Advances in neural information processing systems*, vol. **18**, ed. Y. Weiss, B. Schölkopf & J. Platt, pp. 1291–98. MIT Press. [aFL]

Stocker, A. A. & Simoncelli, E. P. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience* **9**(4):578–85. [VRB]

Suchow, J. W. (2014) *Measuring, monitoring, and maintaining memories in a partially observable mind* (Doctoral dissertation). Harvard University. [aFL]

Suchow, J. W. & Griffiths, T. L. (2016) Deciding to remember: Memory maintenance as a Markov decision process. In: *Proceedings from the 38th annual conference of the Cognitive Science Society*, pp. 2063–68. Cognitive Science Society. [aFL]

Sutherland, S. (2013) *Irrationality: The enemy within*. Pinter & Martin Ltd. [aFL]

Swinnen, S. P. (2002) Intermanual coordination: From behavioural principles to neural-network interactions. *Nature Reviews Neuroscience* **3**:348–59. [ND]

Szollosi, A., Liang, G., Konstantinidis, E., Donkin, C. & Newell, B. R. (2019) Simultaneous underweighting and overestimation of rare events: Unpacking a paradox. *Journal of Experimental Psychology: General* **148**(12):2207–17. Available at: http://dx.doi.org/10.1037/xge0000603. [AS]

Taatgen, N. A. & Anderson, J. R. (2002) Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition* **86**:123–55. [CD]

Taber, C. S. & Lodge, M. (2006) Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* **50**:755–69. [EM]

Tajima, S., Drugowitsch, J. & Pouget, A. (2016) Optimal policy for value-based decision-making. *Nature Communications* **7**:12400–11. doi:10.1038/ncomms12400. [aFL]

Tassinari, H., Hudson, T. E. & Landy, M. S. (2006) Combining priors and noisy visual cues in a rapid pointing task. *The Journal of Neuroscience* **26**(40):10154–63. doi:10.1523/JNEUROSCI.2779-06.2006. [VRB]

Tenenbaum, J. & Griffiths, T. (2001) The rational basis of representativeness. In: *Proceedings from the 23rd annual conference of the Cognitive Science Society*, 84–98. Cognitive Science Society. [aFL]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* **331**:1279–85. [HA]

Theriault, J. E., Young, L. L. & Barrett, L. F. (2019) The sense of should: A biologically-based model of social pressure. *PsyArXiv. Preprint 10.31234/osf.io/x5rbs*. Available at: https://doi.org/10.31234/osf.io/x5rbs. [JET]

Tilman, R. (1991) *Thorstein Veblen and his critics, 1891–1963: Conservative, liberal, and radical perspectives*. Princeton University Press. [HMC]

Todd, P. M. & Brighton, H. (2016) Building the theory of ecological rationality. *Minds and Machines* **26**(1–2):9–30. doi:10.1007/s11023-015-9371-0. [aFL]

Todd, P. M. & Gigerenzer, G. (2012) *Ecological rationality: Intelligence in the world*. Oxford University Press. [aFL]

Todorov, E. (2004) Optimality principles in sensorimotor control. *Nature Neuroscience* **7**(9):907–15. doi:10.1038/nn1309. [aFL]

Todorov, E. & Jordan, M. I. (2002) Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* **5**(11):1226. [WJM]

Tooby, J. & Cosmides, L. (2000) Evolutionary psychology and the emotions. *Handbook of Emotions* **2**:91–115. [KPK]

Tran, R., Vul, E. & Pashler, H. (2017) How effective is incidental learning of the shape of probability distributions? *Royal Society Open Science* **4**(8):170270. [AS]

Treisman, A. M. & Gelade, G. (1980) A feature-integration theory of attention. *Cognitive Psychology* **12**(1):97–136. doi:10.1016/0010-0285(80)90005-5. [aFL]

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M. & Summerfield, C. (2016) Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences* **113**(11):3102–07. doi:10.1073/pnas.1519157113. [aFL]

Tversky, A. & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **5**(2):207–32. doi:10.1016/0010-0285(73)90033-9. [aFL]

Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157):1124–31. doi:10.1126/science.185.4157.1124. [arFL, ANS]

Tversky, A. & Kahneman, D. (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5**(4):297–323. doi:10.1007/BF00122574. [aFL]

van den Berg, R. & Ma, W. J. (2018) A resource-rational theory of set size effects in human visual working memory. *ELife* **7**:e34963. [aFL, WJM]

van der Meer, M., Kurth-Nelson, Z. & Redish, A. D. (2012) Information processing in decision-making systems. *The Neuroscientist* **18**(4):342–59. [rFL]

Van Ravenzwaaij, D., van der Maas, H. L. J. & Wagenmakers, E.-J. (2012) Optimal decision making in neural inhibition models. *Psychological Review* **119**(1):201–15. doi:10.1037/a0026275. [aFL]

Van Rooij, I. (2008) The tractable cognition thesis. *Cognitive Science* **32**(6):939–84. doi:10.1080/03640210801897856. [aFL]

Van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T. & Toni, I. (2011) Intentional communication: Computationally easy or difficult? *Frontiers in Human Neuroscience* **5**:1–18. doi:10.3389/fnhum.2011.00052 [MD]

Varela, F. J., Thompson, E. & Rosch, E. (1991) *The embodied mind: Cognitive science and human experience*. MIT Press. [KPK]

Veblen, T. (1899) *The theory of the leisure class: An economic study of institutions*. Macmillan & Co. [HMC]

Verrecchia, R. E. (1982) Information acquisition in a noisy rational expectations economy. *Econometrica: Journal of the Econometric Society* **50**(6):1415–30. doi:10.2307/1913389. [aFL]

Von Neumann, J. & Morgenstern, O. (1944) *The theory of games and economic behavior*. Princeton University Press. [aFL]

Vourdas, A. (2019) Probabilistic inequalities and measurements in bipartite systems. *Journal of Physics A: Mathematical and Theoretical* **52**:085301. [CM]

Vul, E., Goodman, N. D., Griffiths, T. L. & Tenenbaum, J. B. (2014) One and done? Optimal decisions from very few samples. *Cognitive Science* **38**(4):599–637. doi:10.1111/cogs.12101. [aFL, ANS, WJM]

Vulkan, N. (2000) An economist's perspective on probability matching. *Journal of Economic Surveys* **14**(1):101–18. doi:10.1111/1467-6419.00106. [aFL]

Wald, A. (1950) *Statistical decision functions*. John Wiley & Sons. [ANS]

Waldron, V. R. & Cegala, D. J. (1992) Assessing conversational cognition: Levels of cognitive theory and associated methodological requirements. *Human Communication Research* **18**(4):599–622. doi:10.1111/j.1468-2958.1992.tb00573.x. [MD]

Walleczek, J., ed. (2006) *Self-organized biological dynamics and nonlinear control: Toward understanding complexity, chaos and emergent function in living systems*. Cambridge University Press. [KPK]

Wang, Z., Solloway, T., Shiffrin, R. M. & Busemeyer, J. R. (2014) Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences* **111**(26):9431–36. [HA]

Wang, Z., Wei, X.-X., Stocker, A. A. & Lee, D. D. (2016) Efficient neural codes under metabolic constraints. In: *Advances in neural information processing systems*, vol. **29**, ed. D. D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon & R. Garnett, pp. 4619–27. Curran Associates, Inc. [aFL]

Wason, P. C. (1968) Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20**(3):273–81. doi:10.1080/14640746808400161. [aFL]

Wegner, D. M. & Vallacher, R. R. (1986) Action identification. In: *Handbook of motivation and cognition: Foundations of social behavior*, ed. R. M. Sorrentino & E. T. Higgins, pp. 550–82. Guilford Press. [KP]

Wei, X.-X. & Stocker, A. A. (2015) A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience* **18**(10):1509–17. doi:10.1038/nn.4105. [aFL, EM, WJM]

Wei, X.-X. & Stocker, A. A. (2017) Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences* **114**(38):10244–49. doi:10.1073/pnas.1619153114. [aFL, EM]

Weibel, E. R. (2000) *Symmorphosis: On form and function in shaping life*. Harvard University Press. [JET]

Wells, A. (2011) *Metacognitive therapy for anxiety and depression*. Guilford Press. [EMR]

Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W. & Thomas, M. S. C. (2007) Neuroconstructivism. *Developmental Science* **10**(1):75–83. Available at: https://doi.org/10.1111/j.1467-7687.2007.00567.x. [JET]

Whitton, A. E., Treadway, M. T. & Pizzagalli, D. A. (2015) Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current Opinion in Psychiatry* **28**(1):7–12. Available at: http://doi.org/10.1097/YCO.0000000000000122. [EMR]

Wilson, M. (2002) Six views of embodied cognition. *Psychonomic Bulletin & Review* **9**(4):625–36. [aFL]

Wolfe, J. M. (1994) Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review* **1**(2):202–38. doi:10.3758/BF03200774. [aFL]

Wolpert, D. M. & Ghahramani, Z. (2000) Computational principles of movement neuroscience. *Nature Neuroscience* **3**(11):1212–17. doi:10.1038/81497. [aFL]

Won, I., Gross, S. & Firestone, C. (2019) Impossible somatosensation. *PsyArXiv*. [EM]

Wood, I. (2013) Entrusting western Europe to the Church, 400-750. *Transactions of the Royal Historical Society* **23**:37–73. [HMC]

Woodford, M. (2012) Prospect theory as efficient perceptual distortion. *American Economic Review* **102**(3):41–46. [WJM]

Woodford, M. (2014) Stochastic choice: An optimizing neuroeconomic model. *American Economic Review* **104**(5):495–500. doi:10.1257/aer.104.5.495. [aFL]

Woodford, M. (2016) *Optimal evidence accumulation and stochastic choice* (Technical report). Columbia University. [aFL]

World Health Organization. (2019) *Ten threats to global health*. Available at: https://www.who.int/emergencies/ten-threats-to-global-health-in-2019. [KPK]

Xu, F. & Garcia, V. (2008) Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences* **105**(13):5012–15. doi:10.1073/pnas.0704450105. [VRB]

Yu, Y., Bonawitz, E. & Shafto, P. (2019) Pedagogical questions in parent-child conversations. *Child Development* **90**(1):147–61. [KP]

Yukalov, V. I. & Sornette, D. (2011) Decision theory with prospect interference and entanglement. *Theory and Decision* **70**(3):283–328. [HA]

Zaslavsky, N., Kemp, C., Regier, T. & Tishby, N. (2018) Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* **115**(31):7937–42. doi:10.1073/pnas.1800521115. [aFL, WJM]

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L. & Weintraub, S. (2013) NIH toolbox cognition battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development* **78**(4):16–33. [VRB]

Zénon, A., Solopchuk, O. & Pezzulo, G. (2019) An information-theoretic perspective on the costs of cognition. *Neuropsychologia* **123**(4):5–18. Available at: https://doi.org/10.1016/j.neuropsychologia.2018.09.013. [JET]

Zhang, H. & Maloney, L. T. (2012) Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience* **6**:1. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3261445&tool=pmcentrez&rendertype=abstract [Accessed September 5, 2015]. [DRa]

Zhu, J.-Q., Sanborn, A. N. & Chater, N. (2018a) The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. Available at: https://doi.org/10.31234/osf.io/af9vy. [ANS]

Zhu, J.-Q., Sanborn, A. N. & Chater, N. (2018b) Mental sampling in multimodal representations. In: Advances in neural information processing systems, vol. 31, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett, pp. 5752–63. Curran Associates, Inc. [ANS]

Zipf, G. K. (1949) *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press. [aFL, MD]

https://doi.org/10.1017/S0140525X1900061X Published online by Cambridge University Press