Behavioral and Brain Sciences (forthcoming)

Title

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

Authors

Falk Lieder[1] and Thomas L. Griffiths[2]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany

falk.lieder@tuebingen.mpg.de, https://re.is.mpg.de

[2] Departments of Psychology and Computer Science, Princeton University

tomg@princeton.edu, https://psych.princeton.edu/person/tom-griffiths

Short Abstract

The minds of people, animals, and machines are shaped by the need to quickly solve complex problems with bounded computational resources. Psychologists, economists, neuroscientists, and linguists have begun to leverage this assumption to develop rational models of cognitive strategies and mental representations, and computer scientists have demonstrated that it is possible to derive optimal algorithms for performance-limited hardware. We synthesize these multi-disciplinary advances into an integrative framework and illustrate how it can be used to address classic questions of cognitive psychology, revisit the debate about human rationality, improve the human mind, and connect psychology to neuroscience and artificial intelligence.

Long 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.

Keywords

bounded rationality; cognitive biases; cognitive mechanisms; cognitive modeling; representations; resource-rationality

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 (Rummelhart & McClelland, 1987), Bayesian models of cognition (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), ecological rationality (Todd & Gigerenzer, 2012), and the free-energy principle (Friston, 2012) 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, Laird, & Rogers, 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.

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, Berkes, Orbán, & Lengyel, 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.

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., Gabaix, Laibson, Moloche, & Weinberg, 2006; Dickhaut, Rustichini, & Smith, 2009; 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, Kay, & Khetarpal, 2007; Zaslavsky, Kemp, Regier, & Tishby, 2018; Zipf, 1949), and more recently, psychologists have also begun to incorporate cognitive constraints into rational models (e.g., Griffiths, Lieder, & Goodman, 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 multi-disciplinary 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.

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, Tenenbaum, & Yuille, 2006; Griffiths, et al., 2010; Newell, Shaw, & Simon, 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

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

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, Shaw, & Simon, 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, Slovic, Fischhoff, Layman, & Combs, 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, Griffin, & Kahneman, 2002). In many cases the interpretation of these empirical phenomena as irrational errors has been challenged by subsequent analyses (e.g., Dawes & Mulford, 1996; Hertwig, Pachur, & Kurzenhäuser, 2005; Hahn & Warren, 2009; Gigerenzer, Fiedler, & Olsson, 2012; Gigerenzer, 2015; Fawcett et al., 2014). 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.

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, Todd, & ABC Research Group, 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).

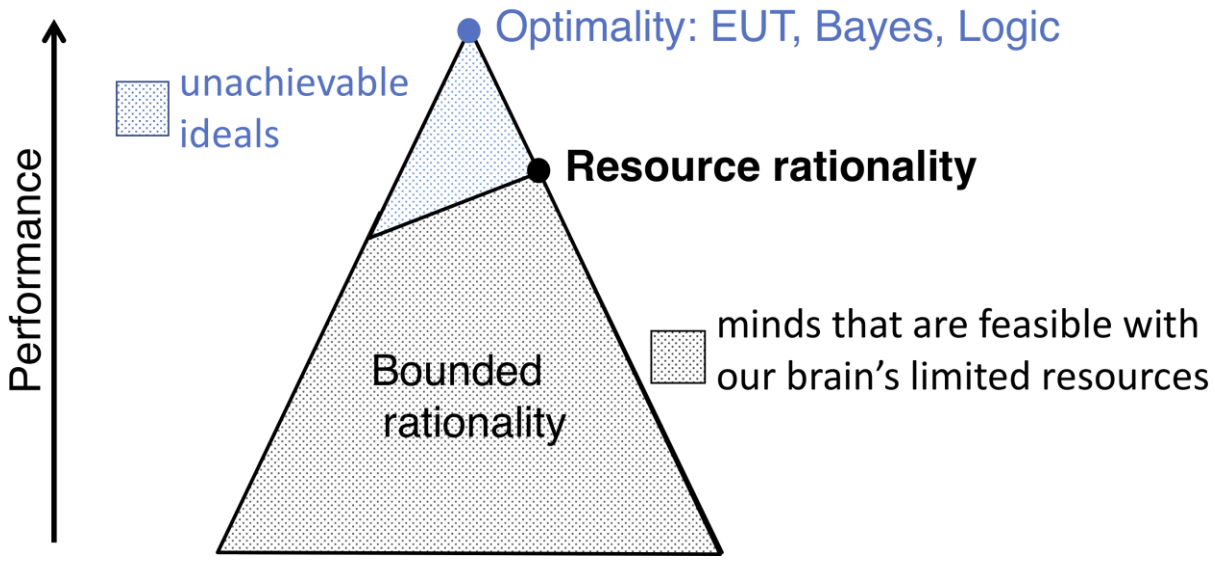

Figure 1: Resource rationality and its relationship to optimality and Tversky and Kahneman's concept of bounded rationality. The horizontal dimension corresponds to alternative cognitive mechanisms that achieve the same level of performance. Each dot represents a possible mind. The gray dots are minds 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 could also 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.[1]

[1] 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.

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. 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 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, Healy, & Hurly, 2002; Shafir, Waite, & Smith, 2002; Latty & Beekman, 2010). 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, 1998a, 1998b; Fawcett, et al., 2014; Johnstone, Dall, & Dukas, 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.

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.

Box 1. The six steps of rational analysis.

Fortunately, AI researchers have already developed a theory of rationality that accounts for limited computational resources (Horvitz, Cooper, & Heckerman, 1989; 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, Howes, & Singh, 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 Figure 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.

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, Horvitz, & Tenenbaum, 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, Goodman, Griffiths, & Tenenbaum, 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, Strack, & Strzalecki, 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, Griffiths, & Goodman, 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-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

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

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 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) takes away from a person's budget for later ones (future episodes). To capture this, let the random variable $\text{cost}(t_h, \rho, \lambda)$ 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

$$h^\star(b_0, B, E) = \underset{h \in H_B}{\text{argmax}}\, \mathbb{E}_{P(\text{result}|s_0,h,E)}[u(\text{result})] - \mathbb{E}_{t_h,\rho,\lambda|h,s_0,B,E}[\text{cost}(t_h, \rho, \lambda)] \quad (3)$$

where H_B is the set of heuristics that brain B can execute and $s_0 = (w_0, b_0)$ comprises the initial state of the external world (w_0) and the person's internal 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 (Houston & McNamara, 1999; McNamara & Weissing, 2010; Dukas, 1998a), 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, Lieder, & Griffiths, 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

$$h^{\star}(b_0, B, I) = \underset{h \in H_B}{\operatorname{argmax}} \mathbb{E}_{E|I}\left[\mathbb{E}_{P(\text{result}|s_0,h,E)}[u(\text{result})] - \mathbb{E}_{t_h,\rho,\lambda|h,s_0,B,E}[\text{cost}(t_h,\rho,\lambda)]\right] \quad (4).$$

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 useful methodology for elucidating what the mind does and why it does what it does (Griffiths, Kemp, & Tenenbaum, 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 their time and cost demands. 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, resource-rational analysis postulates an abstract computational architecture —a set of elementary operations and their costs — with which the mind might solve it. Next, resource-rational analysis derives the optimal algorithm for solving the problem identified at the computational level with

the abstract computational architecture described in Equation 3, thereby pushing the principles of rational analysis toward Marr's algorithmic level (see Figure 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. 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, Lieder, & Goodman, 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.

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, a cost in 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 2).
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.

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.

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; Lieder, Griffiths, & Hsu, 2018) or parametrizing cognitive mechanisms and optimizing the values of those parameters (Lewis et al., 2014), but it is also possible to develop computational methods that discover complex resource-rational cognitive strategies automatically (Callaway, Gul, Krueger, Griffiths, & Lieder, 2018; Callaway, Lieder, Das, Krueger, Gul, & Griffiths, 2017; Lieder, Krueger, & Griffiths, 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.

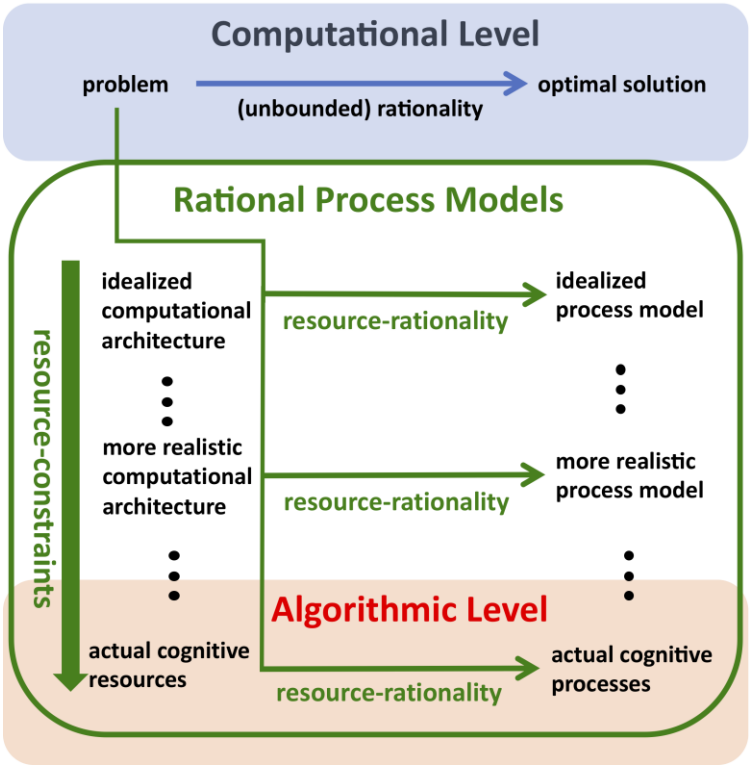

Figure 2. Rational process models can be used to connect the computational level of analysis to the algorithmic level of analysis. The principle of resource-rationality allows us to derive rational process models from assumptions about a system's function and its cognitive constraints.

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 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.

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, Brown, Moehlis, Holmes, & Cohen, 2006; Gabaix et al., 2006; Lieder et al., 2017b; Reis, 2006; Sanjurjo, 2017; C. A. Sims, 2003; Verrecchia, 1982). This cost could include an explicit price that people must pay to purchase information (e.g., Verrechia, 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, Dean, and Martin (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, Laibson, Moloche, and Weinberg (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. In fact, Lieder et al. (2018b) derived that a bounded-optimal decision-maker should prioritize potential outcomes according to the product of their probability and how much their utility depends on the chosen action – thus overweighting scenarios that are extremely good or extremely bad. This model was able to provide a unifying explanation for numerous biases in decisions from experience, frequency estimation, memory recall, and decisions from description.

A refined version of this model can additionally explain how the payoffs' may affect people's risk preferences (Nobandegani, Castanheira, Otto, & Shultz, 2018).

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.

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, Drugowitsch, & Pouget, 2016).

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, Tunney, & McCarthy, 2002; Vulkan, 2000).

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).

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., Stocker, Simoncelli, & Hughes, 2006; Howes, Warren, Farmer, El-Deredy, & Lewis, 2016; Khaw, Li, & Woodford, 2017). However, Beck, Ma, Pitkow, Latham, and Pouget (2012) argue 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, Yadav, & Murawski, 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.

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 et al. (2005, 2006) 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, Gul, Krueger, Griffiths, & Lieder, 2018; Lieder, Callaway, Gul, Krueger, & Griffiths, 2017; Lieder et al., 2017b). Callaway et al. (2018) computed the optimal cognitive strategy for planning based on the opportunity cost of thinking imposed by people's limited time and finite

processing speed by solving Equation 3 exactly and found that the resource-rational heuristic predicted people's planning process substantially better than the myopic model of Gabaix et al. (2006) 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., 2017b), this method rediscovered previously proposed heuristics, such as Take-the-Best (Gigerenzer & Goldstein, 1996), 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.

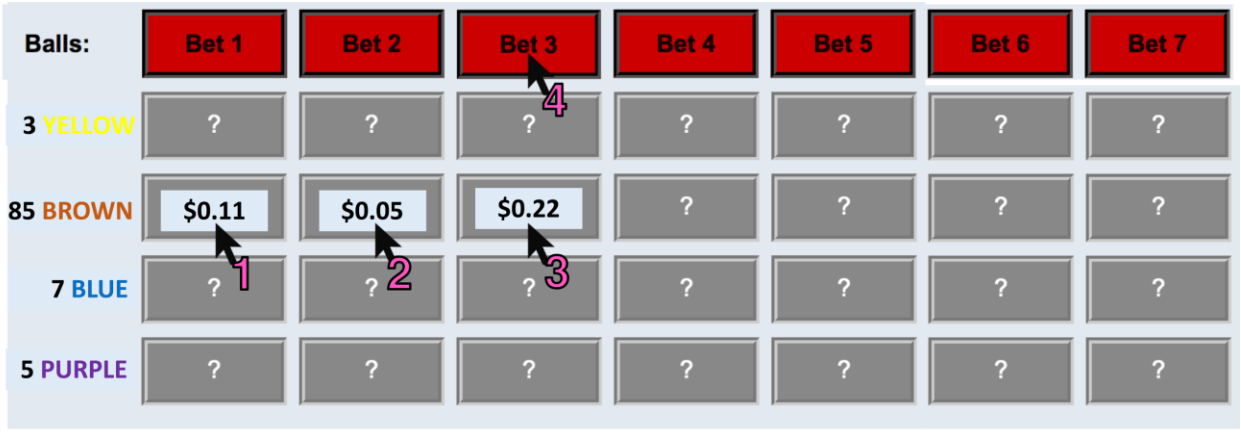

Figure 3: Illustration of the resource-rational SAT-TTB heuristic for multi-alternative risky choice in the Mouselab paradigm where participants choose between bets (red boxes) based on their initially concealed payoffs (gray boxes) for different events (rows) that occur with known probabilities (leftmost 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, Krueger, and Griffiths (2017) derived this strategy as the resource-rational heuristic for low-stakes decisions where one outcome is much more probable than all others.

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, Yadav, & Murawski, 2018; Bossaerts & Murawski, 2017) 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.

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, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 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, Smittenaar, Dolan, & Dayan, 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 sub-problems 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.

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.

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.

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.

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, Duggan, Kalidindi, Tseng, and Lewis (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, Jacobs, & Knill, 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 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.

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, Dean, & Leahy, 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, Oudeyer, Lopes, & Baranes, 2013) or meta-level Markov decision processes (Lieder, Shenhav, Musslick, &

Griffiths, 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. (2017c) 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, 2017). 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 (2017) 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).

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., 2018a, 2018b). 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, Dayan, Griffiths, & Lagnado, 2017), medical diagnoses, and natural scenes (Dasgupta, Schulz, & Gershman, 2017a; Dasgupta, Schulz, Goodman, & Gershman, 2017b, 2017c). 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., 2017b; Dasgupta et al., 2017c).

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, Botvinick, & Cohen, 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, Sokol-Hessner, & Daw, 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., 2017c; Shenhav et al., 2013), how people can flexibly switch between alternative strategies (Lieder & Griffiths, 2017; Payne, Bettman, & Johnson, 1993), and people's occasional lapses of self-control (Boureau et al., 2015).

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. (2018) 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, Chater, & Brown, 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 et al., 2012; C. R. Sims, 2016). 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, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 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.

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, Harb, & Precup, 2017; Botvinick, 2008; Solway et al., 2014) that achieve the best possible trade-off between planning efficiency and reduced behavioral flexibility.

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 et al., 2012; C. R. Sims, 2016; van den Berg et al, 2017). Resource-rational analysis has already led to a fundamental rethinking of the limits of working memory (C. R. Sims, et al. 2012; C. R. Sims, 2016; Van den Berg & Ma, 2017), attention (Van den Berg & Ma, 2017), and cognitive control (Howes, Lewis, & Vera, 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, Sims, Jacobs, & Knill, 2014). Taking this approach even further, van den Berg and Ma (2017) 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, Holmgren, & Juola, 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, Musslick et al., 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, Schwemmer, Gershman, & Cohen, 2014; Musslick et al., 2016) supports the alternative view that capacity limits for multitasking reflect parallel processes competing for limited local resources (Allport, Antonis, & Reynolds, 1972; Meyer & Kieras, 1997a, 1997b; Navon & Gopher, 1979). The bottleneck that the neural pathways of different functions compete for shared representations may itself 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).

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 Figure 4).

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 Figure 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).

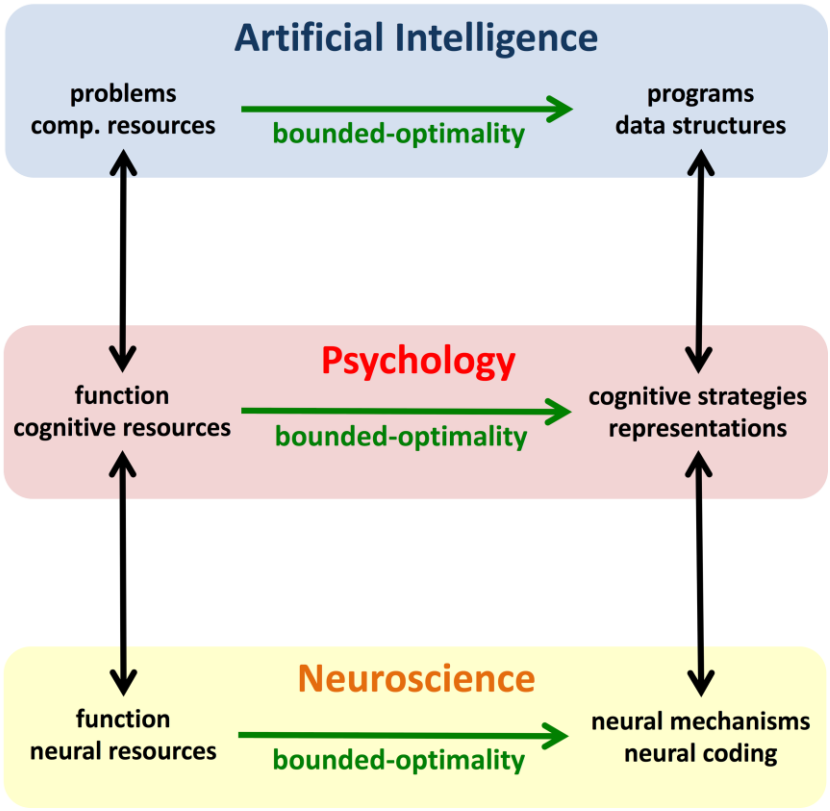

Figure 4: Resource-rational analysis connects levels of analysis.

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, Wei,

Stocker, & Lee, 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, Van der Maas, & Wagenmakers, 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).

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, Griffiths, & Navarro, 2010). Some key AI advances have been inspired by neuroscience and psychology (Hassabis, Kumaran, Summerfield, & Botvinick, 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, Ullman, Tenenbaum, & Gershman, 2017; Nobandegani, 2017).

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., 2018a). Similarly, people's estimates of numerical quantities often being biased toward potentially irrelevant values that were brought to their attention right before they were asked to make their estimate might reflect a rational speed-accuracy trade-off that terminates the process of adjusting the initial estimate when the expected improvement in accuracy drops below the expected time cost of adjustment (Lieder et al., 2018b). Additional cognitive biases that have been shown to be compatible with bounded optimal cognitive mechanisms include the intransitivity of people's

preferences (Tsetsos et al., 2016), 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., 2017a, 2017b), 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, 2009; 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 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 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.

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 (Lieder, Krueger, & Griffiths, 2017; Callaway, Gul, Krueger, Griffiths, & Lieder, 2018) make it possible to discover and teach resource-rational heuristics automatically (Lieder, Callaway, Krueger, Das, Griffiths, Gul, 2018). Teaching optimal heuristics is most appropriate when people perform substantially worse than the resource-rational strategy. In other situations, people's heuristics might already make optimal use of their cognitive resources but the computational complexity of the problem might exceed their cognitive capacities. In cases like this, one might either restructure the environment to simplify the computational problems it poses or augment cognitive capacity. One example of the former is reframing probabilistic reasoning problems in terms of natural frequencies rather than conditional probabilities (Gigerenzer & Hoffrage, 1995; 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, Chen, & Griffiths, in revision).

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.

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 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. Eq. 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 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 resource-rational analysis is meant to uncover are arbitrary and irrational (Gilovich, Griffin, & Kahneman, 2002; Marcus, 2009; Ariely, 2009) and call for different organizing principles (e.g., Kahneman, 2003) such as evolutionary history (e.g., Buss, 1995; Marcus, 2009; 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, 2009). 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, Gul, Krueger, Griffiths, & Lieder, 2018; Callaway, Gul, Krueger, Griffiths, & Lieder, in prep.) 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.

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. 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 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.

References

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), 225–235. DOI: 10.1080/00335557243000102

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249-277. DOI: 10.1037/0033-295X.85.4.249

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Psychology 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), 1036-1060. DOI:10.1037/0033-295X.111.4.1036

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703-719. DOI: 10.1037/0033-295X.96.4.703

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

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

Ariely, D. (2009). *Predictably Irrational*. New York, NY: Harper Collins.

Atkinson, R. C., Holmgren, Je., & Juola, J. F. (1969). Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics*, *6*(6), 321–326. DOI:10.3758/BF03212784

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

Bacon, P.-L., Harb, J., & Precup, D. (2017). The Option-Critic Architecture. Proceedings from

AAAI-17: The 31st Association for the Advancement of Artificial Intelligence Conference On Artificial Intelligence (San Francisco, CA), 1726–1734.

Bateson, M., Healy, S. D., & Hurly, T. A. (2002). Irrational choices in hummingbird foraging behaviour. *Animal Behaviour, 63*(3), 587-596.

Beck, J., Ma, W., 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

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in cognitive sciences, 4*(3), 91-99.

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

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), 700–765. DOI: 10.1037/0033-295x.113.4.700

Bossaerts, P., & Murawski, C. (2017). Computational Complexity and Human Decision-Making. *Trends in Cognitive Sciences*, *21*(12), 917–929. DOI: 10.1016/j.tics.2017.09.005

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

Botvinick, M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208. DOI: 10.1016/j.tics.2008.02.009

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), 700-710 DOI:10.1016/j.tics.2015.08.013

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

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), 301-338. DOI: 10.1037/rev0000061

Buss, D. M. (1995). Evolutionary psychology: A new paradigm for psychological science. *Psychological inquiry*, *6*(1), 1-30.

Butko, N. J., & Movellan, J. R. (2008). I-POMDP: An infomax model of eye movement. Proceedings from *ICDL 2008: 7th IEEE International Conference on Development and Learning* (Monterey, CA), 139–144. DOI: 10.1109/DEVLRN.2008.4640819

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

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

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

Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, *105*(7), 2183–2203. DOI: 10.3386/w19876

Caplin, A., Dean, M., & Leahy, J. (2017). *Rationally inattentive behavior: Characterizing and*

Generalizing Shannon Entropy. NBER Working Paper No. 23652. Cambridge, MA: National Bureau of Economic Research.

Caplin, A., Dean, M., & Martin, D. (2011). Search and Satisficing. *American Economic Review, 101*(7), 2899-2922. DOI: 10.1257/aer.101.7.2899

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

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

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287–291. DOI: 10.1016/j.tics.2006.05.007

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

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2017). Amortized hypothesis generation. *bioRxiv,* 137190. DOI: 10.1101/137190

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2017). Remembrance of Inferences Past. *bioRxiv*, 231837. DOI: 10.1101/231837

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. DOI: 10.1038/nn1560

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), 201-211.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (1st Edition). Cambridge, MA: The Massachussets Institute of Technology Press.

Dickhaut, J., Rustichini, A., & Smith, V. (2009). A neuroeconomic theory of the decision process. *Proceedings of the National Academy of Sciences*, *106*(52), 22145–22150. DOI:10.1073/pnas.0912500106

Dolan, R., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. DOI:10.1016/j.neuron.2013.09.007

Dukas, R. (Ed.). (1998a). *Cognitive ecology: the evolutionary ecology of information processing and decision making*. University of Chicago Press.

Dukas, R. (1998b). Constraints on Information Processing and Their Effects on Behavior. In R. Dukas (Ed.) *Cognitive ecology: the evolutionary ecology of information processing and decision making*. University of Chicago Press.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, *9*(2), 111–118. DOI: 10.1111/1467-9280.00020

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, *51*(4), 380.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, *30*(4), 447–460. DOI: 10.1177/0146167203261889

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, *59*, 255–278. DOI: 10.1146/annurev.psych.59.103006.093629

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), 153-161.

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), 129–146. DOI: 10.3758/s13415-013-0236-9

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

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), 119–130. DOI: 10.1016/j.tics.2010.01.003

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Z. W. Pylyshyn (Ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, 139-150. Norwood, NJ: Ablex.

Frank, M., & Goodman, N. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998. DOI: 10.1126/science.1218633

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

Friedman, M., & Savage, L. J. (1952). The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy*, *60*(6), 463-474.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127.

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). *Speed, Accuracy, and the Optimal Timing of Choices* (Working paper). Cambridge, MA: Massachusetts Institute of Technology.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, *129*(4), 1661–1710. DOI: 10.1093/qje/qju024

Gabaix, X. (2016). *Behavioral Macroeconomics via Sparse Dynamic Programming.* NBER Working paper No. w21848. Cambridge, MA: National Bureau of Economic Research.

Gabaix, X. (2017). *Behavioral Inattention.* NBER Working Paper No. 24096. Cambridge, MA: National Bureau of Economic Research.

Gabaix, X., & Laibson, D. (2005). *Bounded Rationality and Directed Cognition* (NBER and Harvard working paper). Cambridge, MA: National Bureau of Economic Research.

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, *96*(4), 1043–1068. DOI: 10.1257/aer.96.4.1043

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, *26*(10), 2103–2134. DOI:10.1162/NECO_a_00638

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), 273–278. DOI: 10.1126/science.aac6076

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism". *Review of Philosophy and Psychology*, *6*, pp. 363–383. doi: 10.1007/s13164- 015-0248-1.

Gigerenzer, G., Fiedler, K. and Olsson, H. (2012). Rethinking cognitive biases as environmental consequences. In P. M. Todd, G. Gigerenzer, and the ABC Research Group (Eds.).

Ecological rationality: Intelligence in the world. New York: Oxford University Press. pp. 80–110.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. DOI: 10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669. DOI: 10.1037/0033-295X.103.4.650

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

Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: The Massachusetts Institute of Technology Press.

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. New York, NY: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, UK: Cambridge University Press.

Gittins, J., Glazebrook, K., & Weber, R. (2011). *Multi-Armed Bandit Allocation Indices* (2nd Edition). Chichester, UK: John Wiley & Sons.

Glymour, C. (1987). Android epistemology and the frame problem. In Z. W. Pylyshyn (Ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, 63-75. Norwood, NJ: Ablex.

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), 236–243.

DOI:10.1016/S1364-6613(00)01662-4

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), 585–593. DOI: 10.1016/j.tics.2013.09.001

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), 357–364. DOI: 10.1016/j.tics.2010.05.004

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

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. DOI: 10.1111/tops.12142

Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. Proceedings from *The 23rd Annual Conference of the Cognitive Science Society* (Edinburgh, Scotland), 370-375. Austin, TX: Cognitive Science Society.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. DOI: 10.1111/j.1467-9280.2006.01780.x

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

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), 263–268. DOI:10.1177/0963721412447619

Gul, S., Krueger, P.M., Callaway, F., Griffiths, T.L., & Lieder, F. (2018). Discovering Rational Heuristics for Risky Choice. *The 14th biannual conference of the German Society for Cognitive Science, GK,*

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

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704-732. DOI: 10.1037/0033-295X.114.3.704

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, *116*(2), 454-461. DOI: 10.1037/a0017522

Halpern, J. Y., & Pass, R. (2015). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory*, *156*(C), 246–268. DOI: 10.1016/j.jet.2014.04.007

Harman, G. (2013). Rationality. In H. LaFollette, J. Deigh, & S. Stroud (Eds.), *International Encyclopedia of Ethics*. Hoboken: Blackwell Publishing Ltd.

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.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. DOI: 10.1016/j.neuron.2017.06.011

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. New York, NY: Oxford University Press.

Hedström, P., & Stern, C. (2008). Rational choice and sociology. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd Edition). Basingstoke, UK: Palgrave Macmillan.

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

Hertwig, R., & Hoffrage, U. (2013). *Simple Heuristics In a Social World*. New York, NY: Oxford University Press.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211-237. doi:10.1037/a0025940

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), 258–278. DOI: 10.1111/tops.12084

Horvitz, E. J., Cooper, G. F., & Heckerman, D. E. (1989). Reflection and action under scarce resources: Theoretical principles and empirical study. Proceedings from *IJCAI-89: The 11th International Joint Conference on Artificial Intelligence* (Detroit, Michigan), Volume 2, 1121-1127.

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), 1192-1223. DOI: 10.1111/cogs.12271

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological*

Review, *116*(4), 717-751. DOI: 10.1037/a0017187

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), 368-391. DOI:10.1037/a0039996

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., … Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103. DOI: 10.1073/pnas.1414219112

Icard, T. (2014). Toward boundedly rational analysis. Proceedings from *The 36th Annual Conference of the Cognitive Science Society* (Quebec, Canada), Volume 1, 637–642. Austin, TX: Cognitive Science Society.

Icard, T., & Goodman, N. D. (2015). A Resource-Rational Approach to the Causal Frame Problem. Proceedings from *The 37th Annual Meeting of the Cognitive Science Society* (Pasadena, CA). Austin, TX: Cognitive Science Society.

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. http://doi.org/10.1098/rstb.2002.1063

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York, NY: Cambridge University Press. DOI:10.1017/CBO9780511808098.004

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*, 267–293. New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. DOI: 10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. DOI: 10.2307/1914185

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. DOI: 10.1126/science.1218811

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

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), 12868–12873. DOI: 10.1073/pnas.1609094113

Khaw, M. W., Li, Z., & Woodford, M. (2017). *Risk Aversion as a Perceptual Bias.* NBER Working Paper No. 23294. Cambridge, MA: National Bureau of Economic Research.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. DOI: 10.1016/j.tins.2004.10.007

Knill, D. C., & Richards, W. (1996). *Perception As Bayesian Inference*. Cambridge, MA: Cambridge University Press.

Kool, W., & Botvinick, M. M. (2013). The intrinsic cost of cognitive control. *The Behavioral and Brain Sciences*, *36*(6), 697–698. DOI: 10.1017/S0140525X1300109X

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*,

427(6971), 244-247. DOI: 10.1038/nature02169

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

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160. DOI:10.1016/j.cogsys.2006.07.004

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*, 316–338. Malden, MA: Blackwell Publishing.

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), 307-312.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. DOI:10.1016/S0960-9822(03)00135-0

Levy, W. B., & Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation*, *8*(3), 531–543. DOI: 10.1162/neco.1996.8.3.531

Levy, W. B., & Baxter, R. A. (2002). Energy-Efficient Neuronal Computation via Quantal Synaptic Failures. *Journal of Neuroscience*, *22*(11), 4746–4755. DOI: 20026456

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

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), 551–578.

Lieder, F., Callaway, F., Krueger, P. M., Das, P., Griffiths, T. L., Gul, S. (2018). Discovering and Teaching Optimal Planning Strategies In *The 14th biannual conference of the German Society for Cognitive Science, GK*.

Lieder, F., Chen O. X., & Griffiths, T. L. (under review). Cognitive prostheses for goal achievement.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762-794. DOI: 10.1037/rev0000075

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In P. Bartlett, F. C. N. Pereira, L. Bottou, C. J. C. Burges, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, 2690–2798. Red Hook, NY: Curran Associates, Inc.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). 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., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349. DOI:10.3758/s13423-017-1286-8

Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). An automatic method for discovering rational heuristics for risky choice. Proceedings from *The 39th Annual Conference of the Cognitive Science Society* (London, UK), 2567–2572. Austin, TX: Cognitive Science

Society.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2017). Rational metareasoning and the plasticity of cognitive control. *The Public Library of Science Computational Biology*, *14*(4): e1006043. DOI: 10.1371/journal.pcbi.1006043

Locke, E., & Latham, G. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. DOI: 10.1037/0003-066x.57.9.705

Lohmann, S. (2008). Rational choice and political science. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd Edition). Basingstoke, U.K.: Palgrave Macmillan. DOI: 10.1007/978-1-349-58802-2_1383

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. DOI:10.1016/j.cognition.2012.09.010

Marcus, G. (2009). *Kluge: The Haphazard Evolution of the Human Mind*. Boston, MA: Houghton Mifflin Harcourt.

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA: The Massachusetts Institute of Technology Press.

Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1), 272–298. DOI:10.1257/aer.20130047

McNamara,_J. M., & Weissing, F._J._ (2010). Evolutionary Game Theory. In_T. Székely, A. J. Moore, & J.Komdeur (Eds.). Social Behaviour: Genes, Ecology and Evolution,_88–106,_

Cambridge, UK: Cambridge University Press.

Meyer, D. E., & Kieras, D. E. (1997). 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

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review*, *104*(4), 749-791. DOI: 10.1037//0033-295X.104.4.749

Mill, J. S. (1882). *A System of Logic, Ratiocinative and Inductive* (8th Edition). New York, NY: Harper and Brothers.

Milli, S., Lieder, F., & Griffiths, T. L. (2019). *A Rational Reinterpretation of Dual-Process Theories*, Preprint. DOI:

Milli, S., Lieder, F., & Griffiths, T. L. (2017). When does bounded-optimal metareasoning favor few cognitive systems? Proceedings from *AAAI-17: The 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, Volume 31, 4422-4428. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. Psychological Review, 115(2), 502-517.

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. Proceedings from *The 38th Annual Conference of the Cognitive Science Society* (Philadelphia, PA), 1547–1552.

Austin, TX: Cognitive Science Society.

Musslick, S., Saxe, A. M., Ozcimder, K., Dey, B., Henselman, G., & Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures. Proceedings from *The 39th Cognitive Science Society Conference* (London, UK), 829-834. Austin, TX: Cognitive Science Society.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*(3), 214-255. DOI: 10.1037/0033-295X.86.3.214

Neuman, R., Rafferty, A., & Griffiths, T. (2014). A bounded rationality account of wishful thinking. Proceedings from *The 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166. DOI: 10.1037/h0048495

Newell, A., & Simon, H. A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

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), 1792–1804. DOI:10.1242/jeb.017574

Nobandegani, A. (2017). *The Minimalist Mind: On Minimality in Learning, Reasoning*. Georgetown, Canada: McGill-Queen's University Press.

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. *Computing Research Repository, arXiv Preprint:*1801.09848.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data

selection. *Psychological Review*, *101*(4), 608-631. DOI: 10.1037/0033-295X.101.4.608

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning (Oxford Cognitive Science)*. New York, NY: Oxford University Press.

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), 607-609. DOI:10.1038/381607a0

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. DOI: 10.1016/S0042-6989(97)00169-7

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. DOI: 10.1016/j.conb.2004.07.007

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), 164–170. DOI:10.1177/0963721414529144

Pashler, H. E., & Sutherland, S. (1998). *The Psychology of Attention* (Volume 15). Cambridge, MA: Massachusetts Institute of Technology Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge, UK: Cambridge University Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. DOI:10.1073/pnas.1012551108

Polania, R., Woodford, M., & Ruff, C. (2018). Efficient coding of subjective value. *biorXiv*, 358317. DOI: 10.1101/358317

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108. DOI:10.1037/0033-295X.85.2.59

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441. DOI:10.1073/pnas.0610341104

Reis, R. (2006). Inattentive consumers. *Journal of Monetary Economics*, *53*(8), 1761–1800. DOI:10.3386/w10883

Rumelhart, D. E., & McClelland, J. L. (1987). *Parallel Distributed Processing* (Volume 1). Cambridge, MA: Massachusetts Institute of Technology Press.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1–2), 57–77. DOI:10.1016/S0004-3702(97)00026-X

Russell, S. J., & Subramanian, D. (1995). Provably Bounded-Optimal Agents. *Journal of Articial Intelligence Research*, *2*(1), 575–609. DOI: 10.1613/jair.133

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. DOI: 10.1016/j.tics.2016.10.003

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. DOI: 10.1037/a0020511

Sanjurjo, A. (2017). Search with multiple attributes: Theory and empirics. *Games and Economic Behavior*, *104*, 535–562. DOI: 10.2139/ssrn.2460129

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

Segev, Y., Musslick, S., Niv, Y., & Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. Proceedings from *The 40th Annual Conference of the Cognitive Science Society* (Madison, WI). Austin, TX: Cognitive Science Society.

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), 180-187.

Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250. DOI:10.1002/bdm.413

Shenhav, A., Botvinick, M. M., & Cohen, J. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. DOI:10.1016/j.neuron.2013.07.007

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

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405-410.

Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research*, *7*(2), 99–111. DOI:10.1086/208799

Siegler, R., & Jenkins, E. A. (1989). *How children discover new strategies*. New York: Psychology Press.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. DOI: 10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. DOI: 10.1037/h0042769

Simon, H. A. (1982). *Models of Bounded Rationality: Emperically Grounded Economic Reason* (Volume 3). Cambridge, MA: Massachusetts Institute of Technology Press.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. DOI: 10.1016/S0304-3932(03)00029-1

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, *96*(2), 158–163. DOI: 10.1257/000282806777212431

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition*, *152*, 181–198. DOI:10.1016/j.cognition.2016.03.020

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807-930. DOI: 10.1037/a0029856

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

Sosis, C., & Bishop, M. (2014). Rationality. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 27–37. DOI: 10.1002/wcs.1263

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York: Oxford University Press.

Sterling, P., & Laughlin, S. (2015). *Principles of Neural Design*. Cambridge, MA: Massachusetts Institute of Technology Press.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*(3736), 652-654. DOI:10.1126/science.153.3736.652

Stewart, N. (2009). Decision by sampling: the role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology, 62*(6), 1041–1062. DOI:10.1080/17470210902747112

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

Stocker, A., Simoncelli, E., & Hughes, H. (2006). Sensory adaptation within a Bayesian framework for perception. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems,* Volume 18, 1291–1298. Cambridge, MA: Massachusetts Institute of Technology Press.

Suchow, J. W. (2014). *Measuring, monitoring, and maintaining memories in a partially observable mind* (Doctoral dissertation). Cambridge, MA: Harvard University.

Suchow, J. W., & Griffiths, T. L. (2016). Deciding to remember: memory maintenance as a Markov decision process. Proceedings from *The 38th Annual Conference of the Cognitive Science Society*, 2063–2068. Austin, TX: Cognitive Science Society.

Sutherland, S. (2013). *Irrationality: The Enemy Within*. London, UK: Pinter & Martin Ltd.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, *7*, 12400-12411. DOI: 10.1038/ncomms12400

Tenenbaum, J., & Griffiths, T. (2001). The Rational Basis of Representativeness. Proceedings from The *23rd Annual Conference of the Cognitive Science Society,* 84–98. Austin, TX: Cognitive Science Society.

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

Todd, P. M., & Gigerenzer, G. (2012). *Ecological Rationality: Intelligence in the World*. New York, NY: Oxford University Press.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907-915. DOI: 10.1038/nn1309

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1*(1), 46–54. DOI: 10.1111/j.1467-9280.1990.tb00067.x

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

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), 3102-3107. DOI: 10.1073/pnas.1519157113

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. DOI: 10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. DOI: 10.1126/science.185.4157.1124

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

Van den Berg, R., & Ma, W. J. (2017). A rational theory of the limitations of working memory and attention. *bioRxiv*, 151365. DOI: 10.1101/151365

Van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2012). Optimal decision

making in neural inhibition models. *Psychological Review*, *119*(1), 201-215. DOI:10.1037/a0026275

Van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*(6), 939–984. DOI:10.1080/03640210801897856

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica: Journal of the Econometric Society*, *50*(6) 1415–1430. DOI: 10.2307/1913389

Von Neumann, J., & Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

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

Vulkan, N. (2000). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, *14*(1), 101–118. DOI: 10.1111/1467-6419.00106

Wang, Z., Wei, X.-X., Stocker, A. A., & Lee, D. D. (2016). Efficient Neural Codes under Metabolic Constraints. In D. D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, *29*, 4619–4627. Red Hook, NY: Curran Associates, Inc.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. DOI: 10.1080/14640746808400161

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625-636.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding

can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509-1517. DOI:10.1038/nn.4105

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249. DOI:10.1073/pnas.1619153114

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. DOI: 10.3758/BF03200774

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11), 1212-1217. DOI: 10.1038/81497

Woodford, M. (2014). Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, *104*(5), 495–500. DOI: 10.1257/aer.104.5.495

Woodford, M. (2016). *Optimal Evidence Accumulation and Stochastic Choice* (Technical report). New York, NY: Columbia University.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1800521115

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least-Effort.* Oxford, England: Addison-Wesley Press.