Bayesian reverse-engineering considered as a research strategy for cognitive science

Carlos Zednik carlos.zednik@ovgu.de
Frank Jäkel fjaekel@uos.de

forthcoming in *Synthese*

Abstract

Bayesian reverse-engineering is a research strategy for developing three-level explanations of behavior and cognition. Starting from a computational-level analysis of behavior and cognition as optimal probabilistic inference, Bayesian reverse-engineers apply numerous tweaks and heuristics to formulate testable hypotheses at the algorithmic and implementational levels. In so doing, they exploit recent technological advances in Bayesian artificial intelligence, machine learning, and statistics, but also consider established principles from cognitive psychology and neuroscience. Although these tweaks and heuristics are highly pragmatic in character and are often deployed unsystematically, Bayesian reverse-engineering avoids several important worries that have been raised about the explanatory credentials of Bayesian cognitive science: the worry that the lower levels of analysis are being ignored altogether; the challenge that the mathematical models being developed are unfalsifiable; and the charge that the terms 'optimal' and 'rational' have lost their customary normative force. But while Bayesian reverse-engineering is therefore a viable and productive research strategy, it is also no fool-proof recipe for explanatory success.

Acknowledgments

The authors would like to thank Cameron Buckner, Tomer Ullman, and Felix Wichmann for comments on an earlier draft of this paper, as well as the anonymous reviewers. A preliminary version of this work was presented at the Annual Conference of the Cognitive Science Society (Zednik & Jäkel 2014), as well as at workshops and colloquia in Berlin, Cortina d'Ampezzo, Leiden, Osnabrück, Rauischholzhausen, and Tilburg.

1. Introduction

Whereas the Bayesian approach in cognitive science is thought to be genuinely revolutionary by some, it is considered fundamentally flawed by others. Most discussions of Bayesian cognitive science focus on *Bayesian rational analysis*, a method for developing *ideal observer models* of behavior and cognition as optimal solutions to probabilistic inference tasks in the environment (Anderson 1991a; Oaksford & Chater 2007). Although these models have proven useful for characterizing behavioral and cognitive phenomena in a wide variety of domains, critics have worried that they fall short as explanations. In particular, Bowers & Davis (2012a; 2012b) argue that ideal observers are little more than unfalsifiable "just-so stories", and Marcus & Davis (2015, p. 542. See also: Marcus & Davis 2013) worry that the terms 'optimal' and 'rational' are used far "too often, in too many different ways" to be illuminating. Most famously perhaps, Jones & Love (2011) argue that the method of Bayesian rational analysis does little more than describe phenomena at David Marr's computational level of analysis (Marr 1982), thereby falling short of explaining those phenomena by specifying the mechanisms responsible for them (See also: Colombo & Hartmann 2015; Danks 2008).

This article aims to shift the debate away from Bayesian rational analysis, and toward the broader research strategy of *Bayesian reverse-engineering*. Increasingly, proponents of the Bayesian approach treat the method of Bayesian rational analysis as little more than a computational-level starting point of a "top-down" research strategy that answers questions at all three of Marr's levels. For this reason, a proper evaluation of the explanatory credentials of Bayesian cognitive science should consider not only ideal observer models that speak to issues at the computational level, but also the methods that are used to formulate empirical hypotheses at the algorithmic and implementational levels. That said, although recent years have seen a proliferation of methods of this kind (See e.g. Chater et al. 2011; Griffiths et al. 2010; Griffiths et al. 2015; Hahn 2014; Knill & Pouget 2004; Ma et al. 2006; Sanborn et al. 2010), it remains largely unclear what these methods actually have in common; it is unclear how Bayesian reverse-engineering actually works. In what follows, therefore, the different methods that have recently been used to explore all three levels of analysis within the Bayesian context will be subsumed under a general framework in which *tweaking strategies* are used to ensure the empirical adequacy of ideal observers at the computational level (Section 2), and *heuristic strategies* are used to formulate testable hypotheses at the algorithmic and implementational levels (Section 4). Along the way, it will be argued that alternative construals of the Bayesian approach that focus solely on the computational level—notably, *Bayesian Fundamentalism* (Jones & Loves 2011) and *Bayesian Instrumentalism* (Colombo & Seriès 2012; Danks 2008)—misrepresent the stated aims and established practices of Bayesian cognitive science (Section 3).

Armed with a detailed conception of Bayesian reverse-engineering, it will be possible to properly evaluate the explanatory credentials of Bayesian cognitive science. This article is cautiously optimistic, responding to the critics but also reigning in the revolutionaries. Specifically, in Section 5 it will be argued that Bayesian reverse-engineering evades the most

prevalent worries about the Bayesian approach: It does in fact go beyond the computational level of analysis, it can withstand concerns about "just-so-stories", and it is unharmed by the fast-and-loose application of terms such as 'optimality' and 'rationality'. At the same time, however, it will here be argued that Bayesian reverse-engineering is no fool-proof recipe for explanatory success. Indeed, this research strategy is highly pragmatic in nature, in the sense that its outcome depends on the rather unprincipled application of tweaks and heuristics. As a consequence, there is no guarantee that any particular episode of Bayesian reverse-engineering will actually succeed, and it seems likely that different Bayesian reverse-engineers will advance different, and possibly conflicting, three-level explanations of behavior and cognition.

Before commencing in earnest, it is worth providing a preliminary answer to a question that will be revisited repeatedly: What is new and distinctively *Bayesian* about the Bayesian approach? Insofar as many other approaches in cognitive science—including Marr's (1982) information-processing approach—rely on the same basic reverse-engineering strategy, Bayesian reverse-engineering is in many ways "business as usual". What makes Bayesian reverse-engineering unique, therefore, is not the reverse-engineering methods being invoked, but the *pragmatic context* in which these techniques are deployed. Many of the tweaks and heuristics that drive Bayesian reverse-engineering exploit mathematical concepts and computational tools developed to understand the nature of Bayesian probabilistic inference in highly idealized mathematical domains. Of course, the idea that human behavior can be compared to a probabilistic inference engine is not new, and statistical concepts and tools have been used in the service of psychological and neuroscientific theorizing for a long time (See e.g. Brunswik 1943; Swets et al. 1961; Tanner 1961; Peterson & Beach 1967). Nevertheless, the novelty and promise of Bayesian reverse-engineering lies in the use of recent computational and technological advances in Bayesian statistical analysis: By relying on these advances, it is now possible to deal with increasingly complex and realistic kinds of probabilistic inference. Insofar as investigators in artificial intelligence, machine learning, and statistics have gained a sophisticated theoretical understanding of probabilistic inference in idealized machines, Bayesian reverse-engineers are motivated by the prospect of exploiting this understanding to explain probabilistic inference in biological organisms.

2. Bayesian rational analysis

2.1 Starting at the top

Theoretical discussions of the Bayesian approach are traditionally framed in terms of David Marr's (1982) seminal account of explanation in cognitive science.[1] On this account,

[1] Although Marr's account may have its detractors (See e.g. Anderson 2015), a detailed discussion of its virtues and vices is beyond the scope of the present article. Moreover, although several other accounts of explanation in cognitive science have been proposed (See e.g. Cummins 1983; Milkowski 2013a; Zednik 2011), none have been similarly influential, especially within the Bayesian context. The present discussion will therefore assume a basic familiarity with Marr's account, and will take its plausibility for granted. Marr's account of explanation in cognitive science will have served its current purpose if it helps illuminate the principles of Bayesian reverse-

explaining a cognitive system's behavior involves analyzing the system at three distinct *levels of analysis*. At each level, researchers ask a particular set of questions about the system being investigated (See also: McClamrock 1991). At the *computational* level, investigators ask questions about what a cognitive system is doing, and why. Whereas what-questions are traditionally answered by specifying a mathematical function that maps a cognitive system's inputs onto its outputs, why-questions are answered by considering the "appropriateness" of a cognitive system's behavior with respect to the "task at hand" (Marr 1982, p. 24. See also: Section 3). At the *algorithmic* level of analysis, in contrast, investigators ask questions about how the system does what it does. These questions are traditionally answered by specifying the individual steps needed to compute or approximate the relevant input-output function. Finally, at the *implementational* level of analysis, investigators are concerned with questions about where (and when) in the brain the relevant computations take place. Answers to questions of this kind may be delivered by identifying individual steps of an algorithm with the activity (over a certain period of time) of particular physical structures such as neurons or neural networks in the brain (For discussion see: Zednik, forthcoming).

On Marr's account, each one of the computational, algorithmic, and implementational levels is necessary to explain, or "completely understand" (Marr 1982, p. 4ff), a cognitive system's behavior. Because investigators are often unable to answer questions at all three levels simultaneously, however, they typically consider one level at a time. Notably, investigators disagree about the best *order* in which to proceed: Whereas proponents of "bottom-up" research strategies contend that knowledge of possible implementations and algorithms should guide our understanding of behavioral and cognitive capacities (See e.g. McClelland et al. 2010), Marr himself favored a "top-down" approach in which what- and why-questions at the computational level are answered first:

> "Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that [...] an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by understanding the mechanism (and the hardware) in which it is embodied." (Marr 1982, p. 27. See also: Dennett 1994)[2]

Proponents of the Bayesian approach share Marr's predilection for starting at the top; Bayesian reverse-engineers begin by answering what-questions and why-questions at the computational level of analysis, and from there proceed to answer how-questions and where-questions at the algorithmic and implementational levels.

---

engineering.

[2] In hindsight, the normative force of this claim appears overstated: Marr could not have predicted the degree to which e.g. connectionist modeling and brain imaging techniques would galvanize bottom-up research strategies which begin by answering questions at the implementational and algorithmic levels. Nevertheless, Marr's considerations might still be viewed as a statement of the unique virtues of working from the top down: Whereas bottom-up strategies face the daunting task of making sense of complex physical structures in the brain, top-down approaches allow researchers to view such physical structures in a particular way, as contributing to the production of a behavioral or cognitive phenomenon that is itself already quite well-understood. Of course, bottom-up strategies are likely to have virtues of their own. Thus, cognitive science can only benefit from a heterogeneous methodological landscape.

2.2. Developing ideal observers

The computational-level starting point of Bayesian reverse-engineering is the method of *Bayesian rational analysis*, the aim of which is the development of empirically adequate *ideal observer models*. This method is invoked in several related modeling frameworks: signal detection theory (Green & Swets 1988), ideal observer analysis (Geisler 1989), pattern inference theory (Kersten & Schrater 2002), rational analysis (Anderson 1991a), probabilistic modeling (Oaksford & Chater 2001), and Bayesian modeling (Griffiths, Kemp, & Tenenbaum 2008). Despite their different names and their use in different domains, these modeling frameworks share two important features. First, they all use probability theory as their principal mathematical tool. Second, they share the methodological assumption that it is useful to compare a cognitive system's actual behavior in a specific environmental situation to *optimal* (or ideal, or rational) performance in the same situation. Thus, the method of Bayesian rational analysis is designed to address two immediate concerns: How to mathematically define a cognitive system's task environment, and how to specify optimal performance within that task environment using the concepts and methods of probability theory.

Many (though perhaps not all) behavioral or cognitive capacities can be viewed as solutions to probabilistic inference tasks in an uncertain environment (Brunswik 1943). For example, perceptual capacities can be viewed as solutions to the problem of identifying the state of the environment (from among a range of alternatives) that is most likely to have caused a particular stimulation. Similarly, action can be viewed as a solution to the problem of selecting the behavioral response (relative to other possible responses) that is most likely to benefit the agent. Insofar as agents cannot be certain about the actual state of their environment, nor about the future consequences of their actions, they should take into account the uncertainty inherent in a particular environment.

The best-developed mathematical tool for dealing with inference under uncertainty is probability theory. Bayesian rational analysis harnesses this tool to formalize the probabilistic structure of a particular environment, and to quantify the uncertainty inherent in a given task.[3] Although such formalization can be daunting for natural environments, it is quite feasible for laboratory environments that fall under the experimenter's control.[4] Consider a categorization task in a simple laboratory experiment in which white bars of

[3] There are other ways to formalize inference under uncertainty. Because few of these are as well-developed and as well-known as standard probability theory, however, they are less widely used in cognitive scientific research.

[4] Sometimes, Bayesian rational analysis is reduced to analyzing evolutionarily relevant environments or tasks. While this approach has been very successful in behavioral ecology (Davies et al. 2012), in cognitive science it is often difficult or impossible to know what the correct 'natural' environment or task may be (though a notable exception may be natural image statistics: Simoncelli 2003). However, rational analysis has also been used very fruitfully in artificial laboratory experiments without any appeal to evolution, and often with very little information about the natural environment (Anderson, 1991a). As these applications are conceptually simpler, we will focus on such cases here. The particular example considered here, bar-categorization in a simplified laboratory environment, is representative of the ones being used in categorization studies that use artificial stimuli and define categories as probability distributions (See e.g Ashby & Gott 1988; Fried & Holyoak 1984).

varying lengths are displayed over a black background. Participants are tasked with reporting, to the best of their ability, whether a bar that is presented belongs to category *A* or to category *B.* The experimenter fixes the probability that a bar belongs to either category at 1/2, and defines bar-length to be normally distributed with a mean of 10cm and standard deviation of 1cm in category *A*, and with a mean of 12cm and standard deviation of 1cm in category *B* (Figure 1A). Formally, the random variable *Y* denotes the hidden category membership with a probability distribution, the *prior* p(*Y*):

$$p(Y=A) = 1/2 \text{ and } p(Y=B) = 1/2.$$

The random variable *X* denotes the length of the presented stimulus with a conditional probability distribution, the *likelihood* p(*X*|*Y*):

$$p(X=x|Y=A) = \text{normal}(x; 10\text{cm}, 1\text{cm}) \text{ and } p(X=x|Y=B) = \text{normal}(x; 12\text{cm}, 1\text{cm}).$$

Note that the two category distributions overlap: It is very uncertain to which category an 11cm-stimulus actually belongs (Figure 1A). Also note that this uncertainty is intrinsic to the probabilistic structure of the environment--participants should take it into account when solving the task.

Once the task environment has been specified, the second concern for Bayesian rational analysis is to specify an optimal solution. Given the prior probability p(*Y*) and the likelihood p(*X*|*Y*), *Bayes' theorem* prescribes that the *posterior* probability of a presented stimulus *X* belonging to category *Y* is:

$$p(Y|X) = p(X|Y) \cdot p(Y) / p(X)$$

Figure 1B shows the posterior probability of stimuli belonging to category *A* or *B* as a function of their observed length *x*. The greater the observed length, the more probable it is that the stimulus belongs to category *B* (the increasing curve in Figure 1B; the decreasing curve is the posterior probability for *A*). For stimuli of length 11cm, the probability that it belongs to either of the two categories is exactly 1/2.

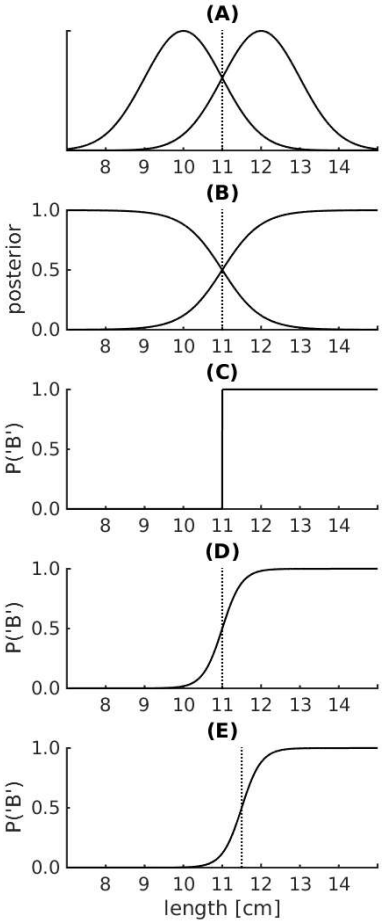

Figure 1: The Bar-Categorization Task

Subjects are presented visually with bars of varying lengths. The bars come from two categories *A* and *B*, one "short" and one "long". The *A*-bars are 10cm long on average but exhibit considerable variance: the left normal distribution in panel (A). The *B*-bars bars are 12cm long on average but also exhibit considerable variance: the right normal distribution in panel (A). As the two category distributions overlap, stimuli around 11cm cannot be assigned to one of the two categories with certainty. Panel (B) depicts the posterior probability of a bar of a certain length to belong to one of the categories under the assumption of equal prior probability. The increasing curve is the posterior probability for category *B* and the decreasing curve for category *A*. Stimuli of 11cm length are equally probable to belong to either category.

The ideal observer categorizes a bar as an *A*-bar whenever it is shorter than 11cm. It is categorized as a *B*-bar otherwise. Panel (C) shows the probability of an ideal observer to respond with B to stimuli of varying lengths. Contrary to the ideal observer's response the responses of (hypothetical) actual subjects, shown in panel (D), are not deterministic: their response probability increases smoothly with stimulus length and exhibits a lot of response variation close to the threshold. Frequently, subjects also have a bias and prefer one response over the other. In panel (E), for example, a subject's response probability is biased to *A*-responses.

Note that the bar-categorization task involves not merely computing a posterior probability, but also acting in a particular way, i.e. *reporting* either '*A*' or '*B*'. In the real world, most actions have consequences that should influence whether or not they will actually be performed. For example, erroneously classifying a poisonous mushroom as edible will be more costly for an organism than erroneously classifying an edible mushroom as poisonous. *Cost functions* codify such relative costs or benefits. In the bar-categorization example, the cost function is simple: Because subjects are instructed to make as few mistakes as possible, mistaking an *A* for a *B* and mistaking a *B* for an *A* incur equal costs, whereas giving the correct answer costs nothing. *Bayesian decision theory* prescribes that, given a posterior probability distribution and a cost function, one should act so as to minimize expected costs. Which response should be given to a stimulus in the bar-categorization task so that the expected number of mistakes is minimized? The answer is intuitive: The optimal stimulus-response function is to give an *A*-response whenever the posterior probability of *A* is higher than the posterior probability of *B*, and a *B*-response otherwise (Green & Swets 1988). Because this is the case whenever a bar is shorter than 11cm, the optimal stimulus-response function has a decision criterion at this exact value (Figure 1B). If a particular stimulus is shorter than this value, the response should be '*A*'; if it is longer, the response should be '*B*' (dotted vertical lines in Figures 1A, 1B). Notably, this stimulus-response function is optimal in the sense that it solves a well-defined mathematical optimization problem in the way prescribed by Bayesian decision theory. Such optimal stimulus-response functions are often called *ideal observer models*[5] (Geisler 1989), and the method of Bayesian rational analysis is designed to deliver models of this kind.[6]

2.3. Tweaking ideal observers to approximate observed behavior

Ideal observer models specify optimal (or "rational") solutions to behavioral and cognitive tasks, construed as probabilistic inference tasks in uncertain environments. What is the relationship between these optimal solutions and actual solutions, i.e. the observable

[5]As the name suggests, ideal observer models were originally used in investigations of visual perception (Geisler 1989). However, the term 'ideal observer' is now also used in contexts that involve other kinds of perception, and even action. We will follow along with this common practice.

[6] A cautionary note on the term 'Bayesian': There is a longstanding debate in philosophy and statistics about the meaning of probability (Hacking 1975). At one end of the debate, *frequentists* take probabilities to be the limits of relative frequencies that are objectively measurable by counting. At the other end, *subjectivists* (often also called 'Bayesians') take probabilities to be expressions of personal beliefs that have to fulfill certain rationality conditions. Although statisticians are becoming less dogmatic about this distinction (See e.g. Efron 2013; Kass 2011), some introductions to the Bayesian approach in cognitive science may lend the impression that it is a defining feature of this approach that probabilities are used to model participants' subjective beliefs. Note, however, that nothing in the method of Bayesian rational analysis as it has been presented thus far is Bayesian in the subjectivist sense (In fact, the ideal observer in the bar-categorization example gives the objectively optimal response in the frequentist sense!). While in philosophy Bayesian epistemology is clearly subjectivist, in Bayesian cognitive science it is not. Probably in order to avoid this confusion, some authors who (according to the present nomenclature) invoke the method of Bayesian rational analysis seem to avoid the label 'Bayesian' for their approaches (See e.g. Geisler 1989; Kersten & Schrater 2002; Anderson 1991a; Oaksford & Chater 2001). That said, even in these research programs, Bayes' theorem plays a central role, and tools from Bayesian statistics and machine learning are routinely invoked to analyze and understand the behavior of ideal observers. For this reason, they too should be subsumed under the Bayesian approach in cognitive science.

behavior of real-world cognitive systems? Proponents of the Bayesian approach have become (in)famous for claiming that human and animal behavior closely approximates the optimal solution of an ideal observer in a wide variety of task environments. An oft-cited example is visual-haptic cue combination (Ernst and Banks 2002), but many other kinds of behavior and cognition have been similarly described as forms of optimal probabilistic inference (For recent reviews see: Berniker & Körding 2011; Pouget et al. 2013; Vilares & Körding 2011. See also Section 5.3).[7]

That said, when an ideal observer model is initially developed, investigators often find discrepancies between the model and observed behavioral data (e.g. Rosas et al. 2005; Rosas et al. 2007). Although an intuitive reaction may be to deny that the observed behavior is optimal, an alternative reaction is to argue that the task environment does not present itself in the same way to the subject as it does to the investigator. That is, the task that the investigator imagined while originally developing the ideal observer model may not be the same as the one the subject is actually trying to solve. In order to better capture a particular body of behavioral data, therefore, the method of Bayesian rational analysis involves *tweaking* the ideal observer model by modifying the underlying assumptions about the task (Step 6 in Anderson 1991a; Step 4 in Griffiths et al. 2015. See also: Swets et al. 1961; Tanner 1961).

Although the term 'tweak' may carry negative connotations of post-hoc data fitting (See discussion in Section 5), here it is meant to highlight the degrees of freedom that researchers may exploit to accommodate the observed behavioral data. Indeed, many different tweaks may often be applied to any given ideal observer, and there are no generally-accepted standards for determining which tweaks to apply when. Thus, the method of Bayesian rational analysis is characterized by a high degree of pragmatism, in which different researchers apply different tweaks in accordance with their preferences and experience (See discussion in Section 4.3). That said, although some tweaks may appear to be ad-hoc, they nevertheless yield formal characterizations that are empirically adequate and predictively powerful. Because of their diversity and pragmatic utility—and because of their role in the eventual outcome of Bayesian reverse-engineering—it is worth considering several of these tweaking methods in detail.

2.3.1. The added-limitations tweak

One common way to tweak an ideal observer model so as to accommodate behavioral data is to add limitations that reflect those of a real-world cognitive system. Consider again the bar-categorization example from above. Figures 1C and 1D show two different psychometric functions—the probability that a subject responds with category *B* as a function of stimulus size. For the ideal observer, the psychometric function is a step function, as in Figure 1C. Although subjects' responses in categorization experiments do sometimes resemble step functions (Ashby & Gott 1988), they usually do not. Instead, the response probability

[7] Strictly speaking, Ernst and Banks (2002) do not use Bayesian arguments when they claim that visual-haptic cue combination is statistically optimal. They use neither priors nor cost functions. However, their maximum likelihood estimator can be given a Bayesian justification under a specific (improper) prior and various cost functions.

changes smoothly, with greater variability closer to the decision criterion (Figure 1D). One reason for this deviation from optimality might be an inability to make fine stimulus discriminations. A subject might easily determine that a 12cm bar is longer than 11cm, but may be unable to reliably determine whether a 11.1cm bar is longer or shorter than 11cm without using a ruler, therefore making mistakes near this boundary. One way to tweak an ideal observer model to account for this kind of variability is to add limitations to the ideal observer's discrimination ability: Just like a physical measurement device, a subject's sensory system is limited by noise. This noise may be added to the likelihood, $p(X|Y)$. By attributing significant proportions of the noise to physical limitations of the receptors or to neuronal firing variability, ideal observer models have become exceedingly useful tools for characterizing subjects' behavior in sensory detection and discrimination tasks (Geisler 1989; Parker & Newsome 1998; Stüttgen et al. 2011). By applying the added-limitations tweak, the observed behavior can be described as being (close to) optimal, under the assumption that the relevant cognitive system is subject to specific quantifiable limitations.

2.3.2. The different-environment tweak

Another way in which actual behavior can deviate from that of the ideal observer is that a subject's decision criterion might be offset from the optimal decision criterion. In Figure 1E, the subject in the bar-categorization task has a response bias toward *A*. In the task environment of the laboratory, categories *A* and *B* are equally probable and all mistakes are equally costly. Hence, in this environment, biased response behavior is clearly not optimal. However, it is conceivable (though not necessarily plausible) that in the real world beyond the laboratory, shorter stimuli are more common than longer stimuli and that $p(A)>p(B)$, in which case a bias toward *A* would in fact be appropriate. As subjects are adapted to the real world, their behavior might be optimal with regard to natural environments, rather than to the laboratory. In this vein, Yang and Purves (2003) measured the distribution of distances of objects to an observer in everyday scenes using a laser scanner and found that some well-known illusions of visual space that occur in the laboratory can be conceptualized as optimal adaptations to the environment. Ideally, a researcher will specify in advance what the natural environment for a particular task is, but in practice there are considerable degrees of freedom in deciding which (aspects of an) environment a subject might be adapted to, and these degrees of freedom can be exploited to tweak an ideal observer's behavior so as to accommodate a particular body of behavioral data.

2.3.3. The subjective-optimality tweak

Yet another reason for a response bias toward *A* (as in Fig 1E) might be that the subject wrongly assumes that category *A* is *a priori* more probable than category *B*. Under this false assumption the observed behavior would be subjectively, albeit not objectively, optimal. Hence, the data could be fit by simply tweaking the ideal observer's prior probabilities without demonstrating that these subjective prior probabilities reflect the probabilities in the natural environment. But tweaking the prior is not the only way to fit the data: It might also be that the subject is not optimizing the cost function imposed on them by the experimenter, but a different one. For example, while the experimenter's task is to minimize

the number of mistakes, subjects might be biased to answer with the left response button (in this case, corresponding to category *A*) just because, due to their left-handedness, this response button is easier to access and therefore incurs a lower cost than the right-hand response button. In other words, a subject's behavior might not be optimal with respect to the cost function imposed by the experimenter, but optimal only with respect to the subject's own subjective costs. Thus, an improved fit to recorded data can also be achieved by tweaking the ideal observer's cost function so that a left-hand response is less costly than that of a right-hand response. Many studies have tweaked priors or cost functions in this way to achieve a better fit to their data, and some specifically set out to measure subjects' subjective priors and costs so as to eventually integrate these into ideal observer models in a variety of task environments (See e.g. Houlsby et al. 2013; Rothkopf & Ballard 2013; Sanborn et al. 2010; Stocker & Simoncelli 2006). Notably, the difference to the different-environment tweak is that no effort is made to demonstrate that the subjective priors and cost-functions are in fact adaptations to some natural environment; instead they are accepted as being purely subjective, and might in fact not reflect the prior and cost-functions corresponding to any particular environment.

2.3.4. The suboptimality tweak

The added-limitations, different-environment, and subjective-optimality tweaks allow researchers to hold on to the postulate that actual behavior is optimal, albeit with respect to different likelihoods, priors, and cost functions than the ones initially assumed by the experimenter. In a sense, these three tweaks aim to remove discrepancies between actual and optimal behavior by modifying the underlying assumptions about the task environment, and about the internal limitations of the subject against which optimality is evaluated. However, it could also be the case that actual behavior is to a certain degree suboptimal, and that this is the reason for the observed discrepancy. Another class of tweaking methods changes the ideal observer model so that it becomes suboptimal itself, e.g. by replacing a deterministic response with a stochastic one, by using an approximation, or by allowing for systematic errors. The difference to the other three tweaks is that no serious effort is made to salvage optimality. In fact, sometimes suboptimality tweaks may be nothing more than ad-hoc methods to link deterministic model predictions with noisy data. Given that there are many potential sources of suboptimality, recent efforts have sought to statistically evaluate their relative contributions in real-world cognitive systems (Acerbi et al. 2014). While a suboptimal response may not be ideal, it can still be good enough for a cognitive system's purposes in a particular task environment (although spelling out what this means can be hard. See: Kwisthout & van Rooij 2013). Most importantly, however, if other tweaking methods are unavailable or undesirable, the suboptimality tweak may be the last (or best) way to accommodate a particular body of behavioral data.

3. Fundamentalism, Instrumentalism, and Reverse-Engineering

The tweaking methods reviewed in Section 2 are designed to eliminate discrepancies between ideal observer models as initially constructed, and the behavioral data being

modeled. That is, they allow proponents of the Bayesian approach to provide empirically adequate descriptions of a cognitive system's behavior. Viewed through the lens of Marr's account of explanation in cognitive science, these tweaks are central to an investigator's ability to answer what-questions and why-questions at the computational level of analysis. Recall that a what-question can be answered by specifying a mathematical function that maps a cognitive system's inputs onto its outputs. Because ideal observer models map stimuli onto responses and can be tweaked to accurately reflect a cognitive system's behavior, they answer what-questions in exactly this way. Recall also that why-questions are answered by evaluating the "appropriateness" of a cognitive system's behavior with respect to the "task at hand" (Marr 1982, p. 24). Although there is no general agreement about what exactly such "appropriateness" amounts to (For discussion see: Shagrir 2010), proponents of the Bayesian approach construe it in terms of statistical optimality: A cognitive system's behavior is appropriate if it constitutes an optimal (or nearly optimal) solution to the relevant task, in the sense prescribed by probability theory. Because ideal observer models specify just such a solution, showing that a system's actual solution can be characterized using an ideal observer model amounts to showing that it is in fact appropriate in this sense. Thus, why-questions can be answered in a rather intuitive way: The system behaves as it does *because* that way of behaving is (nearly) optimal.

That the method of Bayesian rational analysis is designed to answer what- and why-questions at the computational level of analysis is widely acknowledged in the literature (See e.g. Chater et al. 2006; Danks & Eberhardt 2009; Griffiths et al. 2010; Jones & Love 2011; Oaksford & Chater 2007). Nevertheless, it remains a matter of significant controversy whether, and if so how, answering either or both of these kinds of questions is sufficient for the purposes of *explaining* a cognitive system's behavior. In an influential target article, Matthew Jones and Bradley Love (2011) characterize proponents of the Bayesian approach as *Bayesian Fundamentalists,* attributing to them the view that behavior and cognition can be explained "at the computational level [...] without recourse to mechanistic (i.e. algorithmic or implementational) levels of explanation" (Jones & Love 2011, p. 175). On the Fundamentalist construal, the practices described in Section 2 are thought to be sufficient for scientific explanation; additional considerations, including considerations at the algorithmic and implementational levels of analysis, are "essentially irrelevant to understanding cognition" (ibid.).

Bayesian Fundamentalism has been a convenient foil for critics of the Bayesian approach. Jones & Love themselves argue that rejecting the explanatory relevance of the algorithmic and implementational levels "raises the danger of pushing the field of psychology back toward the sort of restrictive state experienced during the strict Behaviorist era" (Jones & Love 2011, p. 176), in which descriptions of cognitive and neural mechanisms were deemed irrelevant to the explanation of behavior. As troublesome as this association with the behaviorist program may be, however, Bayesian Fundamentalism is likely to be a straw man. In the open peer commentary on Jones & Love's target article, Chater et al. argue that Bayesian Fundamentalism is "purely a construct of [Jones & Love's] imagination" (Chater et

al. 2011, p. 194).[8] Indeed, although proponents of the Bayesian approach regularly highlight the computational-level insights being delivered, it is hard to find explicit rejections of the other levels in Marr's tripartite scheme. Instead, they far more commonly express an agnostic attitude toward the algorithmic and implementational levels:

> "the very fact that much cognitive processing is naturally interpreted as uncertain inference immediately highlights the relevance of probabilistic methods at the computational level. This level of analysis is focussed entirely on the nature of the problem being solved—there is no commitment concerning how the cognitive system actually attempts to solve (or approximately to solve) the problem." (Chater et al. 2006, p. 290. Similar statements also occur in: Anderson 1991a, p. 471; Griffiths et al. 2010, p. 362)

An alternative to Bayesian Fundamentalism which better accommodates this agnostic attitude is *Bayesian Instrumentalism* (Colombo & Seriès 2012; Danks 2008). Whereas Fundamentalists deny that questions at the algorithmic and implementational levels are explanatorily relevant, Instrumentalists merely deny that the method of Bayesian rational analysis is designed to answer questions at levels below the computational. That is, although this method might be suitable for characterizing behavior and cognition as a form of (near-) optimal probabilistic inference, its use entails no commitments about underlying mechanisms.

Although Bayesian Instrumentalism may be better able to accommodate statements such as the one by Chater et al., it does not foresee that Bayesian rational analysis is sufficient for genuine explanation. In their articulation of the Instrumentalist view, Colombo & Seriès (2012, p. 9) argue that ideal observer models are "useful instruments, heuristic devices, or tools we employ to predict observable phenomena … or to summarize and systematize data". Although such instruments play an undeniably important role in scientific research, it is doubtful that this role is a genuinely explanatory one; describing a phenomenon is not the same as explaining it, even if the relevant description affords predictive power in actual and counterfactual circumstances (Salmon 1989). In a more recent discussion, Colombo & Hartmann (2015) consider the additional proposal that Bayesian rational analysis contributes to a kind of explanatory unification, because it produces models of a similar mathematical structure for a wide variety of behavioral and cognitive phenomena (See e.g. Griffiths et al. 2010; Tenenbaum et al. 2011). However, it seems unlikely that such unification is any more sufficient for the purposes of cognitive scientific explanation than description or prediction: Insofar as such explanations should "reveal aspects of the causal structure of the mechanism that produces the phenomenon to be explained" (Colombo & Hartmann 2015, p. 15), the purely *mathematical* unification afforded by the method of Bayesian rational analysis has no genuine explanatory force.

These considerations suggest that the ability to answer what-questions in an empirically adequate, predictively powerful and potentially unifying way is insufficient for the purposes

[8] Of course, demonstrating that Bayesian Fundamentalism is in fact a straw man requires more than mere words. Section 4, in which the strategy of Bayesian reverse-engineering is described in detail, shows that many proponents of the Bayesian approach do in fact formulate testable hypotheses not only at the computational level of analysis, but also at the algorithmic and implementational levels. On the Fundamentalist construal, this practice is difficult to accommodate.

of genuine explanation. But what about the fact that the method of Bayesian rational analysis can also be used to answer why-questions? Indeed, many proponents of the Bayesian approach have taken this method's ability to answer why-questions as one of its principal selling-points, and have suggested that doing so is tantamount to providing a *teleological* explanation of the relevant behavioral or cognitive phenomenon (See e.g. Griffiths et al. 2012a; Oaksford & Chater 2007). Here too, however, there are reasons to be skeptical. For instance, Danks (2008) has previously argued that ideal observer models are in fact insufficient even for teleological explanation, because although they might show that some particular behavior is optimal, "there are many other reasons why [it] might occur. People might act optimally because of historical accident, or because there are no other options, or a number of other reasons" (Danks 2008, p. 62). That is, it is a mistake to take literally the claim that a cognitive system behaves as it does *because* that behavior is optimal, as long as it is unclear that this optimality actually played a causal role in phylogenetic or ontogenetic development. Given these considerations, it is unclear that the answers that are being given to why-questions are sufficient for the purposes of teleological explanation, let alone explanation of any other kind.

These considerations suggest that, although Bayesian Instrumentalism may well capture statements that express an agnostic attitude toward the algorithmic and implementational levels, it does not admit of a genuinely explanatory role for the computational-level method of Bayesian rational analysis. Of course, from the perspective of Marr's account of explanation in cognitive science, this is not surprising: Although Marr claimed that the computational level of analysis is "critically important", he also argued that all three levels are needed to attain "complete understanding". That said, there is reason to believe that Bayesian Instrumentalism provides an inadequate foundation on which to evaluate the explanatory credentials of Bayesian cognitive science. Although the Instrumentalist view may in fact be appropriate for discussions of Bayesian rational analysis, it is too narrow for discussions of the Bayesian approach as a whole. Indeed, proponents of the Bayesian approach increasingly go beyond Bayesian rational analysis and the computational-level insights it delivers. To wit, although John Anderson echoes Chater et al.'s agnostic attitude when he claims that Bayesian rational analysis "provides an explanation at a level of abstraction above specific mechanistic proposals", he goes on to suggest that this method

"helps define the issues in developing a mechanistic theory" (Anderson 1991a, p. 471). Similarly, Hahn (2014, p. 8) argues that "rational considerations … are, in fact, part of the route to identifying mechanism or process-level constraints in the first place". More programmatically, Griffiths et al. (2010, p. 357) promote a "top-down or function-first approach", and Chater et al. (2011, p. 196) advocate a "top-down research strategy" which pays attention to insights from "a number of mutually constraining levels of explanation".[9]

These statements imply that it is a mistake to lean too heavily on expressions of agnosticism toward the lower levels of analysis: These expressions concern the limited scope of Bayesian rational analysis, not the explanatory aspirations of the Bayesian approach as a whole. For this reason, the rest of the discussion will focus on showing exactly how proponents of the Bayesian approach intend to go beyond the computational level of analysis, and to thereby answer questions at the algorithmic and implementational levels. In particular, Section 4 will outline the research strategy of *Bayesian reverse-engineering*, a "top-down" approach that begins with the computational-level method of Bayesian rational analysis, and from there answers questions at the lower levels of Marr's hierarchy.[10] Only by considering how this research strategy is used to answer questions at the computational, algorithmic, and implementational levels will it be possible to properly evaluate the explanatory credentials of Bayesian cognitive science.

4. Bayesian reverse-engineering as heuristic search

4.1. Bayesian reverse-engineering

In general, reverse-engineering strategies in cognitive science begin by developing computational-level models of the phenomena being explained, and proceed by inferring the likely composition and organization of the relevant mechanism(s) at the algorithmic and implementational levels.[11] That is, these strategies aim to descend what Daniel Dennett (1987, p. 227) has called a "triumphant cascade through Marr's levels". *Bayesian* reverse-engineering is a research strategy of this kind, the starting point of which is a (tweaked) ideal observer model of the phenomenon being explained. Although the preceding

[9] Some previous discussions of Bayesian Instrumentalism similarly acknowledge the need to go beyond the computational level of analysis. In particular, whereas Danks (2008) criticizes the explanatory value of Bayesian rational analysis insofar as it is confined to the computational level of analysis, he proposes a solution in which considerations of rationality and optimality are also deployed at lower levels. Similarly, upon Colombo & Hartmann's (2015) critique of the explanatory force of (mathematical) unification at the computational level, they propose to consider "what sorts of constraints can Bayesian unification place on causal-mechanical explanation" (Colombo & Hartmann 2015, p. 15). The latter proposal in particular is very much in line with the reverse-engineering view outlined below. That said, whereas Colombo & Hartmann do well to identify constraints that are approximately equivalent to the *push-down* and *unification* heuristics outlined in Section 4, that section will show that the constraints that are in fact imposed on the lower levels are far more numerous, heterogeneous, and unprincipled than previous commentators appear to have recognized.

[10] Notably, Tenenbaum et al. (2011, p. 1279), Chater et al. (2011, p. 196), and Frank (2013, p. 417) each describe their own approach as an exercise in "reverse-engineering the mind".

[11] Milkowski (2013b) provides an alternative account of reverse-engineering in cognitive science which is distinct, but not incompatible with the one sketched here.

discussion already suggests that proponents of the Bayesian approach mean to invoke this research strategy, many important details remain unspecified: When and how can an ideal observer be used to infer the structure and function of psychological processes and neurobiological mechanisms? How can answers to what-questions and why-questions at the computational level be used to answer how-questions and where-questions at the algorithmic and implementational levels?

A helpful backdrop against which to illuminate the principles of Bayesian reverse-engineering is Herbert Simon's account of scientific discovery (Simon et al. 1981)—the process by which scientists develop, evaluate, and refine theories, models, and explanations. On Simon's account, scientific discovery is a form of problem solving in which investigators are tasked with searching a *hypothesis space* of possible solutions. Over this backdrop, Bayesian reverse-engineering can be viewed as a multi-step scientific-discovery problem that is solved by searching three distinct hypothesis spaces—one for each level of analysis (Figure 2). Developing and systematically tweaking an ideal observer model in the ways discussed in Section 2 is the first step, revealing one way to analyze a cognitive system's behavior at the computational level of analysis. The second step is to select one algorithm from a space of possible algorithms for computing or approximating the ideal observer's behavior. Finally, the third step is to choose one implementation from a space of possible implementations of the chosen algorithm. Unlike the first step, the second and third steps of Bayesian reverse-engineering are rarely discussed in the literature (But see: Zednik & Jäkel 2014).

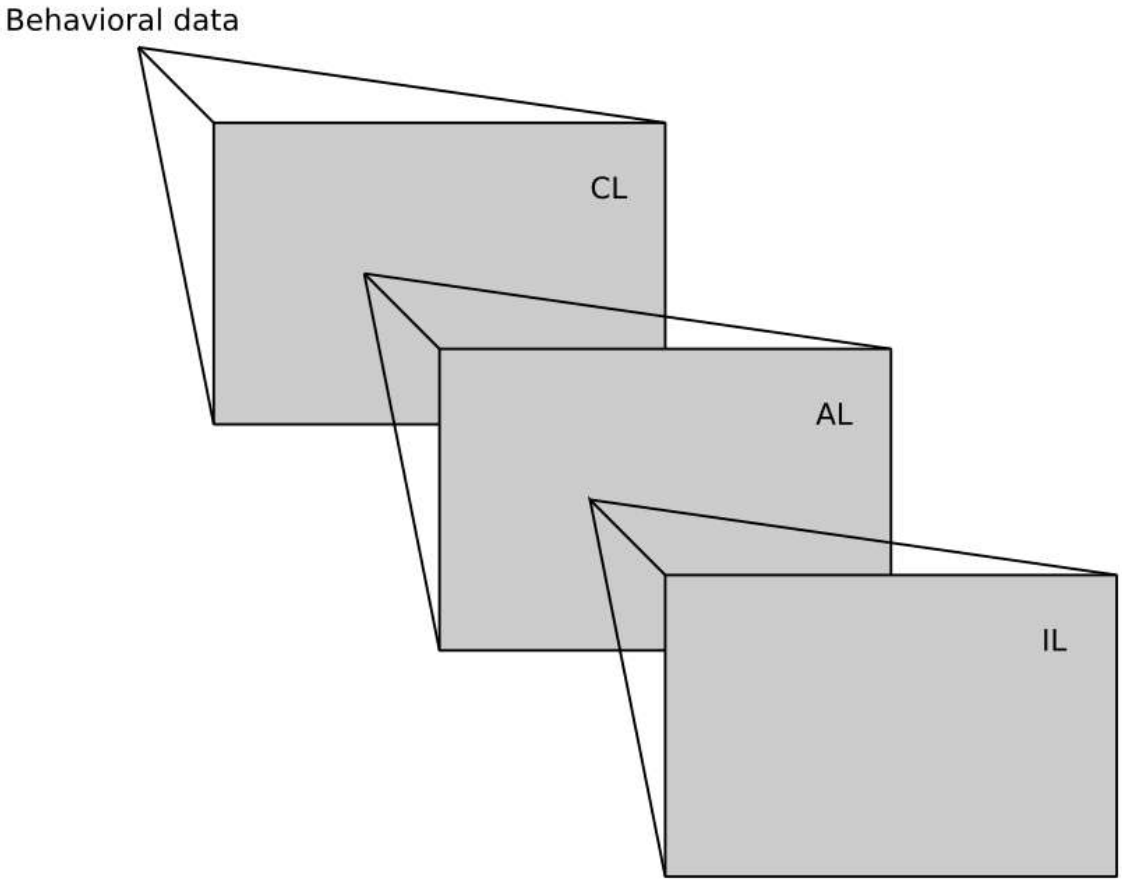

Figure 2: Bayesian reverse-engineering.

Boxes represent search spaces at each of Marr's three levels: computational (CL), algorithmic (AL), and implementational (IL). The starting point for Bayesian reverse-engineering is the method of Bayesian rational analysis, the aim of which is to develop and an ideal observer model, and to tweak it so as to capture a particular body of behavioral data. Different ways to tweak the ideal observer correspond to different points on the CL-space. Given an empirically-adequate ideal observer model, strategies such as the tools-to-theories and push-down heuristics (see below) are invoked to select an algorithm from within the AL-space. As before, different strategies are likely to select different points within the space. Given a particular algorithm, other heuristic strategies such as the possible-implementations heuristic are used to select a possible implementation of that algorithm in neural hardware.

There is reason to worry that suitable algorithmic- and implementational-level solutions will be hard to identify. A general obstacle to scientific discovery as conceived by Simon is the fact that many problems have large hypothesis spaces that cannot be searched exhaustively. In the context of Bayesian reverse-engineering, this obstacle is manifested in the fact that any number of algorithms can often be used to compute a particular function, and that each one of these algorithms might be implemented in many different ways. Whereas investigations at the algorithmic and implementational levels must show how and where the brain computes a particular stimulus-response function, there are many ways in

which this might be the case (See also: Anderson 1978). For this reason, Bayesian reverse-engineers are tasked with exploring exceedingly large hypothesis-spaces at both the algorithmic and implementational levels of analysis.

In order to overcome the obstacle posed by large hypothesis-spaces, Simon appeals to the use of *heuristic strategies* that allow researchers to limit their search to particular regions or points within a particular space. Much as the method of Bayesian rational analysis is characterized by the use of tweaking methods to explore the computational-level search space, the research strategy of Bayesian reverse-engineering is characterized by the use of heuristic strategies that facilitate the search for algorithms and implementations. Therefore, in order to understand how Bayesian reverse-engineering works, as well as to evaluate its likelihood of explanatory success, it is necessary to consider some of these strategies in detail.

4.2. Heuristic strategies for descending the Marrian cascade

4.2.1. The push-down heuristic

One of the most widely-used heuristics for Bayesian reverse-engineering might be called the *push-down heuristic* (cf. Griffiths et al. 2012b; 2015). Bayesian reverse-engineers invoke this strategy whenever they "push" a particular ideal observer's mathematical structure from the computational level onto the algorithmic level of analysis. In order to understand exactly what this "pushing" amounts to, and therefore, which algorithmic-level hypotheses are likely to be selected, it is necessary to distinguish between *implicit* and *explicit* ways of characterizing an ideal observer's mathematical structure. Consider again the ideal observer model for the bar-categorization task, which categorizes *A*-bars and *B*-bars in an optimal way. To characterize this ideal observer implicitly is just to state that it is the function, i.e. the stimulus-response mapping, that minimizes the expected number of mistakes. In contrast, to characterize the ideal observer explicitly is to specify an algorithm that computes this function. The implicit characterization is important for understanding the problem that the ideal observer is supposed to solve, thereby answering a why-question at the computational level. However, it is also necessary to answer the what-question: In order to explicitly characterize the behavior of the ideal observer, it is also necessary to specify an algorithm for computing it.

Notably, a single *generic algorithm* can be used to explicitly characterize many different ideal observers: Compute posterior probabilities using Bayes' theorem, use these to compute the expected cost for all possible actions, and from those choose the action that minimizes the expected cost. While this generic algorithm stands out as a particularly common way of explicitly characterizing ideal observers, there are many others. For example, in the case of the bar-categorization task, the very same ideal observer model (i.e. the very same optimal stimulus-response function) can also be computed by an algorithm that applies a simple rule: Output '*A*' if the input is shorter than 11cm, and '*B*' otherwise. Although this algorithm merely applies a simple decision criterion, it behaves *as if* it applies Bayes' theorem, computes posterior probabilities and expected costs, and chooses an

optimal action according to the principles of Bayesian decision theory. The input-output behavior of the two algorithms is indistinguishable. Because at the computational level it only matters *what* is computed and not *how*, both algorithms explicitly characterize the same ideal observer.

The algorithms that are used by investigators to explicitly characterize an ideal observer (at the computational level of analysis) should not be confused with those that are used by the brain (viewed at the algorithmic level of analysis) in the production of behavior. Nevertheless, the push-down heuristic serves well to select specific algorithmic-level hypotheses by "pushing" the former onto the latter: Bayesian reverse-engineers invoke this heuristic whenever they consider an algorithm that was previously used to explicitly characterize an ideal observer model of a cognitive or behavioral phenomenon as a possible description of the underlying processes that contribute to this phenomenon (See also: Colombo & Hartmann 2015; Colombo & Seriès 2012; Griffiths et al. 2012b; 2015).

The push-down heuristic is most clearly invoked in recent discussions of the *Bayesian Coding Hypothesis*, in which the generic algorithm for characterizing an ideal observer is "pushed" onto the algorithmic level. To wit, the Bayesian Coding Hypothesis is an algorithmic-level hypothesis which claims that "the brain represents information probabilistically, by coding and computing with probability density functions or approximations to probability density functions" (Knill & Pouget 2004, p. 713). Notably, this hypothesis is often supplemented with complementary proposals at the implementational level. For example, recent studies have sought to identify the location of probabilistic representations in the brain (Vilares et al. 2012), and to identify the neural traces of Bayesian computation (Berkes et al. 2011; Ma et al. 2006; Ostwald et al. 2012). In this way, the generic algorithm for characterizing ideal observers at the computational level quite directly guides investigations at the algorithmic and implementational levels.

Although the push-down heuristic is often used to motivate the Bayesian Coding Hypothesis, it can also motivate the consideration of alternative hypotheses. Indeed, which particular hypothesis is selected depends on the particular way in which ideal observers are characterized at the computational level; different explicit characterizations will lead, by way of the push-down heuristic, to the formulation of different algorithmic-level hypotheses. Remember that in the bar-categorization example there were at least two ways in which to explicitly characterize the ideal observer. Given a characterization in terms of the generic algorithm, the push-down heuristic highlights the Bayesian Coding Hypothesis. In contrast, given a characterization in terms of an algorithm that merely checks whether the length of the stimulus is greater than a certain decision criterion (11cm), the push-down heuristic highlights what might be called the *Decision Criterion Hypothesis*: Subjects compare the perceived stimulus length against a particular criterion. Unlike the Bayesian Coding Hypothesis, the Decision Criterion Hypothesis does not postulate that subjects actually represent probabilities or compute over them using Bayes' theorem.[12] In this sense,

[12] In fact, the idea of a decision criterion for solving probabilistic categorization tasks has inspired the development of several learning algorithms that are all inconsistent with the Bayesian Coding Hypothesis, because they do not depend on the representation of probability distributions (Dorfman & Biderman 1971; Kac 1969; Stüttgen et al. 2013; Thomas 1973).

what "comes out" at the algorithmic level when applying the push-down heuristic depends greatly on what "goes in" at the computational level.[13]

The fact that different computational-level analyses lead to the consideration of different algorithmic-level hypotheses suggests that it is very important to be clear about what the push-down heuristic allows Bayesian reverse-engineers to do. By applying this heuristic, Bayesian reverse-engineers are able to formulate some particular algorithmic-level hypothesis, and to seek evidence for or against that hypothesis through additional behavioral and neuroscientific investigation.[14] But the push-down heuristic cannot itself provide evidence for any particular hypothesis over an empirically equivalent alternative. Because every ideal observer model can be explicitly characterized in many different ways, and because each way of characterizing will, via the push-down heuristic, lead to the consideration of different algorithmic-level hypotheses, there is no way of knowing *a priori* which one of these hypotheses is most likely to be true. Notably, although this fact has already been recognized by many (See e.g. Griffiths et al. 2010, p. 362; Clark 2013, p. 191; Colombo & Seriès 2012, p. 10-11), it has not been recognized by all:

> "Recent psychophysical experiments indicate that humans perform near-optimal Bayesian inference in a wide variety of tasks, ranging from cue integration to decision making to motor control. *This implies that* neurons both represent probability distributions and combine those distributions according to a close approximation to Bayes' rule." (Ma et al. 2006, p. 1, emphasis added)

The view expressed in this statement—what might be called *Bayesian Realism* (Colombo & Seriès 2012)—posits that an ideal observer model's empirical adequacy constitutes empirical evidence for the claim that the algorithm used to explicitly characterize this model is actually implemented in the brain. This view is untenable, however: Many different algorithms can be used to compute the very same ideal observer, each one of which would recommend, by way of the push-down heuristic, a different algorithmic-level analysis (See also: Maloney & Mamassian 2009). In summary, although the push-down heuristic can be

[13] Notably, what "comes out" at the algorithmic level might itself also feed back on what "goes in" at the computational level. Specifically, the intention to later invoke the push-down heuristic may already influence the selection of the computational-level tweaks discussed in Section 2.3: Investigators might tweak an ideal observer in one way rather than another just because that way suggests, via the push-down heuristic, certain candidates at the algorithmic level. In particular, it seems very natural to use the added-limitations tweak together with the push-down heuristic because the limitations can be set up to map directly onto hypothesized mechanisms. In so doing, investigators change *what* the ideal observer does and simultaneously select a corresponding hypothesis about *how* the relevant mechanism works. Importantly, although the push-down heuristic thus establishes an intimate link between the computational and algorithmic levels, this does not mean that there exists a level "between the computational and the algorithmic" (Griffiths et al. 2015). On the questions-based interpretation of Marr's framework outlined in Section 2, appeals to in-between levels are confusing: The computational level of analysis concerns what- and why-questions; the algorithmic level concerns how-questions; what kinds of questions occupy the space in between? (See also: Zednik, forthcoming)

[14] For example, if investigators seek to test the Bayesian Coding Hypothesis they will often manipulate the basic building blocks of the generic algorithm, namely likelihoods, priors, and cost functions, and attempt to predict a subject's performance under these manipulations (e.g. Battaglia et al. 2013; Houlsby et al. 2013; Maloney & Mamassian 2009). They may also search for neural correlates of likelihoods, priors, and cost functions (e.g. Berkes et al. 2011; Vilares et al. 2012). Other hypotheses will be tested differently.

used by Bayesian reverse-engineers to formulate specific algorithmic-level hypotheses, it does not license the inference that any particular hypothesis is actually true.

4.2.2. The tools-to-theories heuristic

Whereas the push-down heuristic encourages Bayesian reverse-engineers to co-opt algorithms from the computational level, other heuristics encourage them to introduce algorithms from completely different domains of inquiry. One such heuristic has elsewhere been called the *tools-to-theories heuristic* (Gigerenzer 1991). In general, researchers invoking this heuristic assume that the mechanism responsible for some phenomenon resembles an instrument, tool, or analytic technique that has previously been used to measure, study, or describe that phenomenon. Historically, signal detection theory and ideal observer analysis—two early progenitors of Bayesian rational analysis—were strongly influenced by engineering solutions to detection-problems in radar and sonar technology, as well as by progress in statistical signal processing (Swets 2010). These tools were then adapted to measure, study, and describe human behavior in sensory detection tasks (Green & Swets 1988). More recently, the tools-to-theories heuristic is most clearly invoked by Bayesian reverse-engineers who seek inspiration from technological developments in artificial intelligence, machine learning, and statistics. Indeed, they have argued that "new computational methods for efficient Bayesian inference and learning [in statistics and machine learning] have substantially expanded the range of possible hypotheses concerning representations and algorithms in human inference and learning" (Chater et al. 2011, p. 195), and that "the best algorithms for approximating probabilistic inference in computer science and statistics" can often be considered "candidate models of cognitive and neural processes" (Griffiths et al. 2012b, p. 264).

Consider a recent example due to Sanborn et al. (2010). They compare the relative merits of three distinct algorithms—*Gibbs sampling*, *particle filtering,* and an *iterative algorithm*—for supplementing John Anderson's Bayesian rational analysis of categorization (Anderson 1991b). Whereas the iterative algorithm is an adaptation of "a type of iterative algorithm that has appeared in the artificial intelligence literature" (Anderson 1991b, p. 412), the Gibbs sampling and particle filtering algorithms both belong to the class of *Monte Carlo* algorithms developed in machine learning to approximate Bayesian inference (Andrieu et al. 2003). By comparing the relative performance of these three algorithms in a wide range of experimental paradigms, Sanborn et al. collect considerable evidence, most notably in the form of order effects, to support the hypothesis that human categorization is performed using a particle filtering algorithm.[15] Thus, although this algorithm was originally developed as a tool in machine learning and statistics, the tools-to-theories heuristic has allowed

[15] As these three algorithms only *approximate* the ideal observer, selecting them as algorithmic-level hypotheses simultaneously invokes the suboptimality tweak at the computational-level. Each of the three different approximations makes a concrete proposal for answering the how-question at the algorithmic level. But since they do not compute exactly the same function, they also show a difference in the observable behavior and therefore each answers the what-question at the computational level slightly differently (e.g. by showing different order effects). It is, however, still the case that they answer the why-question at the computational level in the same way, because they are all considered approximations to the untweaked ideal observer (But see: Kwisthout & van Rooij 2013).

Bayesian reverse-engineers to demonstrate that a reasonably adapted variant also happens to be a good description of the psychological processes that contribute to human categorization.[16]

Recall that the outcome of the push-down heuristic depends on the particular way in which an ideal observer is characterized explicitly. In much the same way, the tools-to-theories heuristic highlights many different (and possibly incompatible) algorithmic-level hypotheses, depending on which areas of artificial intelligence, machine learning, and statistics are actually being considered. Indeed, neither one of these heuristics is meant to suggest that any single algorithmic-level hypothesis is in fact true. Rather, they are both designed to facilitate the formulation of such hypotheses, and to thereby make possible their eventual (dis)confirmation through subsequent psychological or neuroscientific research.

4.2.3. The unification heuristic

Although many heuristic strategies drive Bayesian reverse-engineering by facilitating the formulation of candidate hypotheses, other strategies are used to choose between existing alternatives. For example, many Bayesian reverse-engineers invoke the *unification heuristic,* which highlights those algorithmic-level hypotheses that seem most likely to complement not only the ideal observer model for a single behavioral or cognitive phenomenon, but also the ideal observer models for other phenomena. Notably, this heuristic strategy appears to have contributed to the influence of the Bayesian Coding Hypothesis. Whereas the Bayesian Coding Hypothesis is very general and potentially applies to many different task environments beyond the one of the familiar bar-categorization example, the Decision Criterion Hypothesis works well only in that particular task environment—it will be suboptimal, or not at all applicable, in most others. The general applicability of the Bayesian Coding Hypothesis might be viewed as a reason to prefer it over the Decision Criterion Hypothesis, because it has the potential to unify many different phenomena under a common algorithmic theme (See also: Colombo & Hartmann 2015; Ma et al. 2006). Indeed, it seems likely that whereas many proponents of the Bayesian Coding Hypothesis inadvertently commit themselves to the strong but untenable position of Bayesian Realism, they actually mean to do nothing else than combine the push-down and unification heuristics in an effort to reverse-engineer the mind.

4.2.4. The plausible-algorithms heuristic

It might be worried that none of the heuristic strategies considered thus far are sensitive to traditional psychological considerations. For example, known limitations in working memory or attention that have been studied extensively by cognitive psychologists play no role in the selection of algorithms by way of the push-down, tools-to-theories, and unification heuristics. Nevertheless, Bayesian reverse-engineers rarely apply these heuristics blindly. For example, Sanborn et al. (2010) consider Monte Carlo sampling algorithms not only because they are useful tools for approximating optimal Bayesian inference, but also

[16] Another class of algorithms used for the same purpose, *variational inference* (Beal 2003), has proven similarly useful for the purposes of Bayesian reverse-engineering (e.g. Friston 2008; Sanborn & Silva 2013).

because they have certain properties, like working incrementally and giving rise to order effects, that make them psychologically plausible candidates. The *plausible-algorithms heuristic* is invoked whenever available knowledge about cognitive processes is used to guide the selection of candidate hypotheses from the algorithmic-level search space.[17]

Although the plausible-algorithms heuristic is often combined with other heuristics (e.g. push-down or tools-to-theories), it can also be used on its own. For example, exemplar models are a class of well-established algorithms for categorization that have been tested extensively in the literature (See e.g. Nosofsky 1986; Kruschke 1992). As these models can also be used to estimate probability distributions (Ashby & Alfonso-Reese 1995), they have recently been proposed as plausible algorithms for tasks in which human behavior can be described as a form of optimal probabilistic inference (Shi et al. 2010). In this way, algorithms that are known to be plausible in other psychological domains can be co-opted for use in the Bayesian context.[18] Although these algorithms may not explicitly invoke Bayes' theorem or probability distributions, they can be invoked by proponents of the Bayesian approach insofar as they serve to compute or approximate an empirically adequate ideal observer.

4.2.5. The possible-implementations heuristic

The push-down, unification, tools-to-theories, and plausible-algorithms heuristics allow Bayesian reverse-engineers to descend from the computational to the algorithmic level of analysis. Other strategies are used to descend further down the Marrian cascade. Consider what might be called the *possible-implementations heuristic*. Researchers invoke this strategy whenever they consider known principles of brain function or organization to generate hypotheses about possible implementations for a particular algorithm. Although little may be known about the mechanisms responsible for a specific behavioral or cognitive phenomenon, much is known about the general ways in which structures in the brain compute (See e.g. Dayan & Abbott 2001). This knowledge can be exploited to formulate hypotheses about how the brain *might* implement a particular algorithm. For example, it is relatively easy to come up with neurally-plausible proposals for how the brain might represent probability distributions as posited by the Bayesian Coding Hypothesis: The firing rate of a single neuron could directly code log-probabilities; a population of neurons with differing tuning curves may code a probability distribution by a basis function expansion; or the activity of pools of neurons might represent samples from a distribution (Pouget et al.

[17] Psychological considerations may also already enter at the computational level through the added-limitations tweak discussed in Section 2.3.1. Take the bar-categorization example: Adding noise to the stimulus representation of the decision-criterion algorithm is plausible because it is known that subjects' discrimination ability is limited.

[18] Although Bayesian reverse-engineering is a "top-down" research strategy in the sense discussed previously, regular use of the plausible-algorithms heuristic shows how it might be combined with "bottom-up" approaches that start with cognitive or neural principles. In particular, one might seek to determine exactly how an established cognitive architecture (such as ACT-R) could bring about an ideal observer's optimal performance (Cooper & Peebles 2015; Thomson & Lebiere 2013). In general, being a proponent of top-down research strategies does not entail a rejection of bottom-up strategies. In this context, it is also worthwhile to remember that the same John Anderson who developed rational analysis also developed the ACT-R architecture and that the 'R' stands for 'rational'.

2013). Similarly, depending on how probabilities are represented, different neural implementations of Bayesian updating suggest themselves; computing posterior probabilities might be as straightforward as summing the activities of presynaptic neurons in a single postsynaptic neuron (Ma et al. 2006).

Although Bayesian reverse-engineers have invoked the possible-implementations heuristic mostly to search the implementational-level search space below the Bayesian Coding Hypothesis, it can also be applied to other algorithmic-level hypotheses, e.g. the Decision Criterion Hypothesis. Notably, the possible-implementations heuristic can be applied to any algorithmic-level hypothesis, and it does not matter which heuristics were used to select this algorithm in the first place. Although such top-down proposals for the possible implementation of a specific algorithm remain highly speculative, they can still be used to focus neuroscientific research on specific regions of the implementational-level search space. A particularly vivid example is the recent proposal that spontaneous neural activity in the absence of sensory stimulation can be interpreted as a neural signature of Monte Carlo sampling (Berkes et al. 2011; Fiser et al. 2010).[19]

4.3. Will Bayesian reverse-engineering succeed?

Bayesian reverse-engineers are likely to deploy many additional heuristic strategies beyond the ones reviewed here. Moreover, as was already the case for the computational-level tweaking methods discussed in Section 2, there are no generally-accepted standards for determining how and when to apply particular heuristic strategies for the purposes of descending the Marrian cascade. Whereas some earlier discussions may lend the impression that there is a single tried-and-true "recipe" for Bayesian cognitive science (See e.g. Anderson 1991a; Griffiths et al. 2015), Bayesian reverse-engineering is in fact a highly pragmatic research strategy, the course of which depends on researchers' individual background and preferences. Nevertheless, it is worthwhile to consider its likelihood of explanatory success--that is, the likelihood that Bayesian reverse-engineering will yield approximately true (or well-confirmed) hypotheses at the algorithmic and implementational levels of analysis.

The explanatory success of Bayesian reverse-engineering is likely to depend on another aspect of Simon's account of scientific discovery: the sense in which heuristics for scientific discovery are highly *efficient*, but also *fallible* and *systematically biased* (Simon 1996. See also: Wimsatt 1985). As has already been suggested above, most of the heuristics that contribute to Bayesian reverse-engineering serve to formulate testable hypotheses, but not to directly support the claim that any one of these hypotheses is actually true. Although it might be worried that these heuristics therefore do little to *solve* reverse-engineering

[19] Just like the relation between the computational and the algorithmic level (See footnotes 13 and 18), the relation between algorithmic and implementational level is also not completely "top down". Hypotheses at the algorithmic level are often chosen *because* they are easy to map onto neural structures and processes. In the case of Monte Carlo sampling the algorithm definitely preceded the possible implementation (Fiser et al. 2010). However, in the case of log-probabilities the ease with which Bayesian updating could be implemented in neurons is very likely to have influenced the choice of representation at the algorithmic level (Ma et al. 2006).

problems in cognitive science, it is important to consider the difficulties involved in formulating testable hypotheses in everyday scientific practice. Coming up with such hypotheses is extremely difficult, and scientists rarely, if ever, possess a complete, well-articulated list of hypotheses from which they are merely tasked with choosing the best. For this reason, at least as much time and effort goes into the process of formulating testable hypotheses as goes into the process of choosing between them. Although some of the heuristics that drive Bayesian reverse-engineering may do little with respect to the latter, they do greatly facilitate the former; they are catalysts for inspiration, and in this sense make an invaluable contribution to scientific discovery.

At the same time, the reliance on heuristics bears uncertainties and even risks. All of the heuristics introduced above are fallible: Nothing guarantees that the hypotheses highlighted by the push-down, tools-to-theories, unification, plausible-algorithms, possible-implementations, or any other heuristic strategy are actually true. How then is the use of such heuristics an improvement over random guesswork? The importance of this question is amplified if one considers the efficiency outlined just above: If a particular heuristic leads to the formulation of many false hypotheses, it is likely to do more harm than good, because it will lead to the disproportionate consumption of time and scientific resources. In what sense are the heuristic strategies that drive Bayesian reverse-engineering not just efficient, but efficient guides to truth?

This question can be answered by considering the third feature of heuristic strategies: their systematic bias. Most heuristics do not highlight solutions at random, but systematically, by selecting only those solutions that exhibit a particular set of characteristics. The extent to which a heuristic strategy is an efficient guide to truth may depend on the nature of its bias, i.e. the kinds of considerations that are invoked to select individual solutions. The heuristics that drive Bayesian reverse-engineering fall into two broad categories: Those whose systematic bias is *theoretical*, i.e. that invoke considerations rooted in the principles of a particular background theory,[20] and those whose systematic bias is *pragmatic*, i.e. that invoke considerations rooted in the particular set of tools and concepts at a researcher's disposal, as well as in the social structures and institutions in which that researcher is embedded.

The plausible-algorithms and possible-implementations heuristics both exhibit a theoretical bias: They are designed to select just those algorithmic- and implementational-level hypotheses that accord with the established principles of psychology and neuroscience. Insofar as these principles are at least approximately true, the plausible-algorithms and plausible-implementations heuristics can be viewed as reasonable guides to truth; their potential to lead researchers astray is no worse than the fallibility of the background theory whose principles they exploit.

[20] For example, the assumption that many biological systems are modular (and thus, that scientific discovery problems in biology should be solved by focusing on modular solutions) might be justified by appealing to the principle from evolutionary theory that modular systems are more robust, and thus more likely to survive and reproduce, as compared to non-modular systems (Simon 1996).

In contrast, most of the other heuristic strategies outlined above exhibit a distinctively pragmatic bias. Rather than select algorithmic- and implementational-level hypotheses by considering a relevant background theory, heuristics such as push-down and tools-to-theories appeal to the *pragmatic context* in which Bayesian reverse-engineering unfolds: The mathematical and computational tools, concepts, and methods of probability theory and Bayesian decision theory, as well as the social structures and institutions in which researchers working on the nature of probabilistic inference are embedded. As has already been demonstrated above, the push-down heuristic highlights just those algorithmic-level hypotheses that reflect some particular way of characterizing an ideal observer model at the computational level of analysis. Had this characterization been different, either because it invoked a different algorithm to explicitly characterize the ideal observer, or indeed, because it invoked a wholly different body of mathematical concepts and tools, the push-down heuristic would have selected a different algorithmic-level hypothesis. In a similarly pragmatic way, the tools-to-theories heuristic works by selecting just those algorithmic-level hypotheses that correspond to available algorithms in artificial intelligence, machine learning, and statistics. Insofar as investigators working in these disciplines have for years been programming machines to solve the same (or similar) kinds of problems that are being solved by human beings, it is not surprising that Bayesian reverse-engineers find inspiration in these researchers' methods and results.[21] As a consequence, the nature and strength of the interdisciplinary collaborations in which any particular Bayesian reverse-engineer is engaged is likely to greatly influence the outcome and productivity of his or her own research.

Interestingly, similarly pragmatic research strategies have worked well in the past. Although the use of reverse-engineering principles is only recently being applied in the modern Bayesian context, these principles are far from new in cognitive science as a whole (See e.g. Dennett 1994; Milkowski 2013b). Indeed, many of the same heuristics that drive Bayesian reverse-engineering may also play a role in other flavors of reverse-engineering, contextualized by other mathematical concepts and tools (See e.g. Jäkel et al. 2009).[22] For this reason, the current excitement surrounding Bayesian reverse-engineering may have less to do with the novelty of the basic research strategy itself, and more to do with the impression that recent mathematical and technological advances in Bayesian methods can now be used to study behavior and cognition. In fact, whereas classical ideal observer

[21] This point is widely recognized in the literature. For example, Tenenbaum et al. (2011, p. 1279) argue that "What has come to be known as the 'Bayesian' or 'probabilistic' approach to reverse-engineering the mind has been heavily influenced by the engineering successes of Bayesian artificial intelligence and machine learning over the past two decades". In recognition of this influence, many introductions to Bayesian rational analysis and Bayesian reverse-engineering focus explicitly on the presentation of specific mathematical methods, computational tools, and on the interdisciplinary utility of Bayesian statistical inference (See e.g. Griffiths et al. 2008).

[22] As has already been suggested, the novelty of Bayesian reverse-engineering lies not in the basic methodological principles being invoked, but only in the novel use of specific mathematical concepts and methods. Specifically, Bayesian reverse-engineering is uniquely "Bayesian" in exactly two ways. First, its starting point is the method of Bayesian rational analysis, which aims to describe different kinds of behavior and cognition as solutions to problems of probabilistic inference. Second, the solutions Bayesian reverse-engineers are most likely to discover are those that reflect the solutions being found in the interdisciplinary research community of Bayesian artificial intelligence, machine learning, and statistics.

analysis—the earliest progenitors of modern Bayesian reverse-engineering—as used in signal detection theory is among the most successful methods in all of psychology and cognitive neuroscience, it was mostly limited to the exploration of relatively simple detection and discrimination tasks. Insofar as Bayesian reverse-engineering invokes many of the same methodological principles but additionally exploits recent technological developments for solving ever more complex inference problems, it is easy to think that the early successes of ideal observer analysis and signal detection theory may now be replicated in increasingly higher domains.

Whereas there is reason to be optimistic, therefore, it is important not to confuse methodological promise with genuine theoretical progress (See also: Jones & Love 2011). Indeed, many important theoretical questions remain unanswered: Which general principles govern the functioning of the mind and brain? Which architectural features do they exhibit? As a research strategy, the principles of Bayesian reverse-engineering do not themselves prescribe answers to these questions; Bayesian reverse-engineering is not itself an explanation of the mind and brain, but a methodological framework for *developing* explanations. Although it is possible that this strategy will eventually lead to a unified conception of the "Bayesian brain" in the sense that the mechanisms underlying a wide variety of behavioral and cognitive phenomena exhibit similar structural and functional properties (as is suggested by e.g. proponents of the Bayesian Coding Hypothesis), it seems more likely that the diversity of methods and tools in Bayesian artificial intelligence, machine learning, and statistics will contribute to a theoretical conception of the mind and brain as a heterogeneous collection of processes and mechanisms.[23] Given that it is hard to predict which particular tools and concepts will influence Bayesian reverse-engineering in the future, the outcome of this research strategy can only be known by letting it run its course.

5. Of straw men and red herrings

This presentation of Bayesian reverse-engineering provides an overview of the tweaks and heuristics that are used to answer questions at all three levels of analysis, as well as a preliminary evaluation of the chances that this research strategy will eventually succeed. In order to support this evaluation, however, it is also necessary to consider a series of principled worries. As has already been indicated in Section 3, critics have challenged the explanatory credentials of the Bayesian approach. Because discussions of this approach traditionally focus on the computational-level method of Bayesian rational analysis, however, it is unclear to what extent these challenges should also be worrisome to Bayesian reverse-engineers who aim to answer questions all three levels of analysis. Indeed, it

[23] For example, whereas category learning might best be described using particle filters at the algorithmic level (Sanborn et al. 2010), perceptual decision-making might be better described with a decision criterion (Stüttgen et al. 2013), categorical perception with an exemplar model (Shi et al. 2010), and theory learning as stochastic search (Ullman et al. 2012). Even though in each of these domains the method of Bayesian rational analysis was used at the computational level (and in this sense, contributing to a sense of mathematical unification: Colombo & Hartmann 2015), the resulting algorithmic-level models developed by Bayesian reverse-engineers are strikingly different, and possibly even incompatible.

appears that, by focusing on the wider context of Bayesian reverse-engineering, the challenges can be met.

5.1. Lacking mechanisms

One of the most influential critiques of Bayesian cognitive science is Jones & Love's (2011) discussion of Bayesian Fundamentalism, already introduced in Section 3. Recall that Jones & Love worry that the Bayesian approach denies the explanatory relevance of the algorithmic and implementational levels; Bayesian Fundamentalism "eschews mechanism" (Jones & Love 2011, p. 173), and comes dangerously close to behaviorism. Whereas it was already claimed in Section 3 that Bayesian Fundamentalism is a straw man, Section 4 substantiates this claim: Far from denying the explanatory relevance of the algorithmic and implementational levels, Bayesian reverse-engineers are very much willing and able to explore levels below the computational.

There is a kernel of truth in Jones & Love's worry, however. Although Section 3 demonstrates that there is a stated desire to descend the Marrian cascade, and although Section 4 shows how this desire may eventually be satisfied, there continues to be a relative paucity of success stories. Although sophisticated multi-level explanations are available in domains such as visual perception (Geisler 1989; Kersten et al. 2004), perceptual decision making (Gold & Shadlen 2007; Stüttgen et al. 2011), and category learning (Anderson 1991a; Sanborn et al. 2010), there are relatively few examples in other (and especially higher) cognitive domains. That said, even in these higher domains, the desire to pursue a reverse-engineering strategy is apparent and increasingly explicit (See e.g. Frank 2013; Griffiths et al. 2012b; Griffiths et al. 2015; Tenenbaum et al. 2011). On the one hand, therefore, it seems fair to urge that the critics be patient. On the other hand, however, the pragmatic nature of Bayesian reverse-engineering makes it difficult to predict its success.

5.2. Lacking falsifiability

Whereas Jones & Love challenge the Bayesian approach's willingness to go beyond the computational level, other critics question its adequacy at the computational level itself. Recall that this level concerns questions about what a cognitive system is doing, as well as questions about why. Bowers & Davis' (2012a; 2012b) concern lies primarily with the former. As has already been discussed in Sections 2 and 3, proponents of the Bayesian approach answer what-questions by developing and tweaking ideal observers to capture a particular body of behavioral data. Although Bowers & Davis agree that ideal observers can be used to describe "almost any pattern of results" (Bowers & Davis 2012a, p. 394), they worry that not all descriptions are equally explanatory. Indeed, there is an intuitive sense in which answering a question about what a cognitive system is doing is not just a matter of capturing the relevant behavioral data, but also a matter of identifying the task a particular cognitive system is *actually* solving, even if the systems' behavior may be characterized in an empirically adequate way as a solution to several different tasks. Because the method of Bayesian rational analysis involves the use of tweaks, and because there are no general guidelines for determining which tweaks to apply when, different investigators are prone to

answering the same what-question differently,[24] with no principled way of determining which answer is actually correct. On the basis of these considerations, Bowers & Davis conclude that Bayesian rational analysis is little more than a method for developing unfalsifiable "just-so stories" (Bowers & Davis 2012a, p. 410).

Bowers & Davis' worry is troublesome only insofar as there is no way to distinguish good tweaks from bad ones. However, the discussion of Bayesian reverse-engineering in Section 4 suggests that there actually might be such a way. As discussed above, differently-tweaked ideal observers may lead to the formulation of different algorithmic and implementational-level hypotheses, e.g. by way of the push-down heuristic in which the mathematical structure of a particular ideal observer is "pushed down" onto a lower level. Hypotheses at these lower levels are readily falsifiable, by way of the customary evidential criteria of psychology and neuroscience. That is, behavioral experiments and e.g. neuroimaging studies that can be used to determine which algorithms actually contribute to the production of behavior and cognition, and how these algorithms are actually implemented. Insofar as some tweaks may lead to more readily confirmable lower-level hypotheses than others, these tweaks may also be favored over alternatives with an equal degree of empirical adequacy at the computational level. In this sense, falsifiability at the computational level might be "inherited" from falsifiability at the algorithmic and implementational levels.

The general lesson to draw here is that it is a mistake to evaluate the explanatory viability of the Bayesian approach, as Bowers & Davis as well as many others have done, by considering only the computational-level method of Bayesian rational analysis. As the discussion of Bayesian Instrumentalism has already shown, there are reasons to doubt that this method is sufficient for genuine explanation. Nevertheless, because the method of Bayesian rational-analysis is merely the starting point of Bayesian reverse-engineering, it may in fact yield full-fledged scientific explanations that answer questions at all three levels of analysis. Thus, the explanatory credentials of Bayesian cognitive science cannot be determined merely by considering the development and tweaking of ideal observers at the computational level; it is also important to consider the algorithmic- and implementational-level hypotheses that are generated through the application of heuristic strategies. Very much in line with the conception of scientific explanation advanced by Marr, properly evaluating the explanatory credentials of the Bayesian approach involves issues at all three levels of analysis, rather than one.

5.3. Lacking rationality

Worries have also arisen about why-questions. Recall that the method of Bayesian rational analysis can be used to answer why-questions by showing that a cognitive system's behavior is "appropriate" insofar as it matches, or approximates, the optimal or rational behavior of some ideal observer. One worry about the explanatory value of such optimality-demonstrations has already been discussed in Section 3: It is unclear to what extent the fact

[24] For example, if a subject engaged in the bar-categorization task exhibits a sloping psychometric function instead of the optimal threshold function, one researcher might prefer an added-limitations tweak, whereas another applies a suboptimality tweak. Both tweaks allow researchers to capture the same psychometric function by positing very different ideal observers.

that a behavior is optimal actually played a role in phylogenetic or ontogenetic development (Danks 2008). Another worry has recently been articulated by Marcus & Davis (2013; 2015. See also: Bowers & Davis 2012a). On the face of it, an optimality-demonstration appears to be a substantial empirical discovery: A biological organism engages its task environment in accordance with a particular body of arbitrary formal rules, such as the rules of probability theory (See e.g. Griffiths & Tenenbaum 2006; Oaksford & Chater 2001). Indeed, the growing appeal of Bayesian cognitive science stems at least in part from the sense in which optimality-demonstrations of this kind contrast with demonstrations of apparently sub-optimal (or irrational) behavior (See e.g. Tversky & Kahneman 1974). To return to the bar-categorization example from before, whereas a subject's failure to correctly distinguish *A*-bars from *B*-bars would be traditionally interpreted as a failure of human rationality, the tweaks outlined in Section 2 allow proponents of the Bayesian approach to modify priors, likelihoods, or cost functions so as to show that the subject does in fact behave rationally, albeit with regard to these modifications. That said, the significance of such optimality-demonstrations appears greatly diminished by the proliferation and unprincipled application of tweaks: If the specification of the task environment can be tweaked seemingly at will so as to demonstrate that almost any behavior is optimal or rational (But see: Ellsberg 1961; Savage 1972), an optimality-demonstration appears to have lost its normative force (Bowers & Davis 2012a; Marcus & Davis 2013; 2015).

Troublingly, proponents of the Bayesian approach often provide conflicting assessments of the intended normative force of their claims. On the one hand, many important contributions defend the claim that human or animal behavior is genuinely optimal (e.g. Ernst & Banks 2002; Körding & Wolpert 2004), and position themselves against the view that judgment and reasoning are biased and error-prone (e.g. Griffiths & Tenenbaum 2006; Oaksford & Chater 2001). Furthermore, attempts have been made to explain away apparent deviations from optimality by arguing that suboptimal algorithms are in fact "rational under resource constraints" (Gershman et al. 2015; Griffiths et al. 2015; Vul et al. 2014; cf. Love 2015). In contrast, other recent discussions appear to weaken the normative force of optimality-demonstrations by admitting that, for example, "*an* optimal analysis is not *the* optimal analysis for a task" (Goodman et al. 2015, p. 539, emphasis in the original), or that it "[is rarely the goal of Bayesian modeling] to show that people perform optimally at particular tasks" (Griffiths et al. 2012a, p. 412; See also: Frank 2013, footnote 2 and p. 419f). Given these conflicting (and admittedly intransparent) assessments by influential proponents of the Bayesian approach, it is understandable that the critics remain unsatisfied (See: Bowers & Davis 2012b; Marcus & Davis 2015).

The discussion of Bayesian reverse-engineering in Section 4 allows for a very different perspective on these difficulties: Much confusion can be avoided by recognizing that the issue of rationality is in fact a red herring. Proponents of Bayesian reverse-engineering do not need to claim that human or animal behavior is rational in any deep sense, nor do they need to justify appeals to rationality by e.g. invoking evolution. Indeed, tweaks such as the sub-optimality tweak do not assume that cognitive system's are optimal, and the heuristics that are used to explore lower levels of analysis do not make use of rationality considerations at all. Instead, ideal observer models merely provide "a point of reference"

(Peterson & Beach 1967, p. 29) and a "convenient base from which to explore the complex operation of a real organism" (Swets et al. 1961, p. 311, see also Tanner 1961). That is, the notion of rationality need not be used to provide a *sui generis* kind of teleological explanation (Oaksford & Chater 2007; Griffiths et al. 2012a. But cf. Danks 2008; Danks & Eberhardt 2009), but merely to navigate the space of computational-level hypotheses and to guide subsequent investigations at the algorithmic and implementational levels in a way that might even reveal *sub*optimal processes and mechanisms (See e.g. Acerbi et al. 2014; Kruschke 2006; Stüttgen et al. 2013; Sanborn & Silva 2013). For reverse-engineering purposes, whether or not a particular behavioral or cognitive phenomenon is thought to be rational or not is secondary to the ability to uncover the mechanisms responsible for that phenomenon. Disagreements about the explanatory import of optimality-demonstrations, considered in isolation of the wider context of Bayesian reverse-engineering, are moot.

But although considerations of rationality only have heuristic value in the context of Bayesian reverse-engineering, there is no doubt that they are of substantial interest in other contexts. The question of whether humans are in fact rational agents bears on many central debates in evolutionary psychology, education, economics, jurisprudence, politics, and ethics, among others. Given the potential impact psychological investigations of human rationality may have on these debates, it is unsurprising that the apparent optimality-demonstrations of Bayesian cognitive science have received so much attention, and equally unsurprising that Bayesian cognitive scientists themselves often highlight the role of rationality in their works. Nevertheless, as this discussion shows, while questions about rationality are important and interesting in many domains of inquiry, they are a distraction for cognitive scientists interested in reverse-engineering the mind through the "top-down" discovery of mechanisms.

6. Conclusion

As it is characterized here, Bayesian reverse-engineering is a highly pragmatic research strategy. By developing ideal observer models and tweaking them in a variety of different ways, proponents of Bayesian rational analysis describe behavior and cognition at the computational level as a form of near-optimal probabilistic inference. In turn, heuristic search strategies—many of which exploit the pragmatic context defined by a particular body of tools and concepts, but also social structures and institutions—can be used to formulate testable hypotheses at the algorithmic and implementational levels. The widespread use of these strategies shows that proponents of the Bayesian approach are concerned with far more than just the computational level of analysis, and that they are in fact willing and able to discover and describe processes and mechanisms at the algorithmic and implementational levels.

This account of Bayesian reverse-engineering serves to alleviate several worries that have recently arisen about the explanatory credentials of Bayesian cognitive science. While many of these worries appear to be detrimental when directed at the computational-level method of Bayesian rational analysis, they dissolve when considered in the wider context of

Bayesian reverse-engineering. First and foremost, Bayesian reverse-engineers are not Bayesian Fundamentalists, but readily seek to address all three levels of analysis. Second, answers given to questions at the computational level are falsifiable insofar as they facilitate the formulation of testable answers to questions at the algorithmic and implementational levels. Third, at the lower levels it matters little to what extent these answers invoke the notions of 'optimality' or 'rationality'—these notions are used to descend the Marrian cascade, rather than to normatively assess real-world behavior. Fourth and finally, it is important not to confuse the research strategy of Bayesian reverse-engineering for a theoretical framework that advances a specific set of empirical hypotheses about the nature of mind and brain. Indeed, which particular theoretical commitments are actually held by individual Bayesian reverse-engineers is likely to depend on the particular tweaking and heuristic strategies they employ, their experience with and preference for distinct mathematical and computational concepts and tools, and the degree to which they interact with colleagues from disciplines such as artificial intelligence, machine learning, and statistics.

Because it alleviates these worries and provides the methods and tools necessary for developing three-level explanations, Bayesian reverse-engineering is a viable and productive research strategy for cognitive science. That said, it is also a research strategy that is likely to undergo continuous evolution, depending on future technological advances in mathematical and engineering disciplines, as well as on theoretical progress in cognitive psychology and neuroscience. As such, much work remains to be done to determine how best to combine computational-level, algorithmic-level, and implementational-level insights so as to develop integrated three-level explanations of behavior and cognition. As has been shown above, although many different tweaking and heuristic strategies can be used for this purpose, their success is far from guaranteed. For this reason, the explanatory success of Bayesian reverse-engineering can only properly be evaluated by letting this research strategy run its course.

References

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, 20(6), 1-23.

Anderson, B. L. (2015). Can computational goals inform theories of vision? *Topics in Cognitive Science, 7*, 274-286.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review,* 85, 249-277.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471-517.

Anderson, J. R. (1991b). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5-43.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216-233.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*(1), 33-53.

Battaglia, P. W., Hamrick, J., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences USA*, *110*(45), 18327-18332.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD Thesis: The Gatsby Computational Neuroscience Unit, University College London.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*, 83-87.

Berniker, M., & Körding, K. P. (2011). Bayesian approaches to sensory integration for motor control. *WIREs Cognitive Science*, *2*, 419-428.

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, *138*(3), 423–426.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50, 255-272.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences*, *34*(4), 194-196.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287-291.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.

Colombo, M., & Hartmann, S. (2015). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science, [Epub ahead of print].* doi: 10.1093/bjps/axv036

Colombo, M., & Seriès, P. (2012). Bayes in the brain—on Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, *63*(3), 697-723.

Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanations. *Topics in Cognitive Science*, 7, 243-258.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Rational Models of Cognition* (p. 59-75). Oxford University Press.

Danks, D., & Eberhardt, F. (2009). Explaining norms and norms explained. *Behavioral and Brain Sciences, 32*(1), 86-87.

Davies, N. B., Krebs, J. R., & West, S. A. (2012). *An Introduction to Behavioural Ecology*. Chichester, UK: Wiley-Blackwell.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience*. Cambridge, MA: MIT Press.
Dennett, D. (1987). *The Intentional Stance.* Cambridge, MA: MIT Press.
Dennett, D. (1994). Cognitive science as reverse engineering: Several meanings of 'top-down' and 'bottom-up'. In D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), *Logic, Methodology & Philosophy of Science IX* (p. 679-689). Elsevier Science B.V.
Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology*, *8*, 264-284.
Efron, B. (2013). A 250-year argument: belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society*, 50(1), 129-146.
Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643-669.
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.
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.
Frank, M. C. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, *128*, 417-423.
Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 234-257.
Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), 1-24.
Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discrimination. *Psychological Review*, *96*(2), 267-314.
Gershman, S. J.; Horvitz, E. J. & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*, 273-278.
Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254-267.
Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535-574.
Goodman, N., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P., & Hamrick, J. (2015). Relevant and robust. A response to Marcus and Davis (2013). *Psychological Science*, *26*(4), 539-541.
Green, D. M., & Swets, J. A. (1988). *Signal Detection and Psychophysics (reprint edition)*, Los Altos, CA: Peninsula Publishing.
Griffiths, T. L., 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.
Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012a). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(3), 415–422.
Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. 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*, 217-229.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767-773.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012b). Bridging levels of analysis for probabilistic models of cognition. *Psychological Science*, *21*, 263-268.

Hahn, U. (2014). The Bayesian boom: good thing or bad? *Frontiers in Psychology*, 5, 1-12.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge, UK: Cambridge University Press.

Houlsby, N. M. T., Huszár, M. M., Ghassemi, Orbán, G., Wolpert, D. M., & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, *23*, 2169-2175.

Jäkel, F., Wichmann, F. A., & Schölkopf, B. (2009). Does Cognitive Science Need Kernels? *Trends in Cognitive Sciences*, 13, 381-388.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169-231.

Kac, M. (1969). Some mathematical models in science. *Science*, *166*(3906), 695-699.

Kass, R. E. (2011). Statistical inference: the big picture. *Statistical Science*, 26(1), 1-9.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Kersten, D., & Schrater, P. R. (2002). Pattern inference theory: A probabilistic approach to vision. In R. Mausfeld & D. Heyer (Eds.), *Perception and the physical world.* Chichester: John Wiley & Sons, Ltd.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Cognitive Sciences*, *27*(12), 712-719.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-7.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677-699.

Kwisthout, J., & van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research*, 24, 2-8.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*, 240-242.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432-1438.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*, 147-155.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(12), 2351-2360.

Marcus, G. F., & Davis, E. (2015). Still searching for principles: A response to Goodman et al. (2015). *Psychological Science*, *26*(4), 542-544.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines 1*: 185-196.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., & Seidenberg, M. S., (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348-356.

Milkowski, M. (2013a). *Explaining the Computational Mind.* Cambridge, MA: MIT Press.

Milkowski, M. (2013b). Reverse engineering in cognitive science. In Milkowski, M. and Talmont-Kaminski, K. (Eds.), *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental* (p. 12-29). Newcastle upon Tyne: Cambridge Scholars Publishing.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology, 115,* 39-57.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8).

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

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177-188.

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience*, *21*, 227-277.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29-46.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170-1178.

Rothkopf, C. A., & Ballard, D. H. (2013). Modular inverse reinforcement learning for visuomotor behavior. *Biological Cybernetics*, *107*(4), 477-490.

Rosas, P., Wagemans, J., Ernst, M.O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A*, 22(5), 801-809.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, 7(6), 1-12.

Salmon, W. (1989). *Four Decades of Scientific Explanation*. Pittsburgh: Pittsburgh University Press.

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

Sanborn, A., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, *57*, 94-106.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*, 63-106.

Savage, L. J. (1972). *The Foundations of Statistics*. Mineola, NY: Dover. (Original work published 1954)
Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477-500.
Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as mechanisms for performing Bayesian inference. *Psychonomic Bulletin & Review, 17(4)*, 443-464.
Simon, H. A. (1996). *The Sciences of the Artificial (3rd Edition).* Cambridge, MA: MIT Press.
Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese*, *47*(1), 1-27.
Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13, 144-149.
Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578-85.
Stüttgen, M. C., Kasties, N., Lengersdorf, D., Starosta, S., Güntürkün, O., & Jäkel, F. (2013). Suboptimal criterion setting in a perceptual choice task with asymmetric reinforcement. *Behavioral Processes*, *96*, 59-70.
Stüttgen, M. C., Schwarz, C., & Jäkel, F. (2011). Mapping spikes to sensations. *Frontiers in Neuroscience*, *5*(125), 1-17.
Swets, J., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.
Swets, J. A. (2010). *Tulips to thresholds*. Los Altos Hills, CA: Peninsula Publishing.
Tanner, W. P. (1961). Physiological implications of psychophysical data. *Annals of the New York Academy of Sciences*, *89*, 752-765.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.
Thomas, E. A. C. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology*, *10*, 241-264.
Thomson, R. and Lebiere, C. (2013). Constraining Bayesian inference with cognitive architectures: An updated associative learning mechanism in ACT-R. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society.* (p. 318-362). Austin, TX: Cognitive Science Society.
Tversky, A., & Kahneman D. (1974). Judgments under uncertainty. Heuristics and biases. *Science*, 185, 1124-1131.
Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455-480.
Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Körding, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*, 1-8.
Vilares, I., & Körding, K. P. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*.
Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599-637.

Wimsatt, W. C. (1985). Heuristics and the study of human behavior. In D. W. Fiske & R. Shweder (Eds.), *Metatheory in Social Science: Pluralisms and Subjectivities* (p. 293–314). Chicago, IL: University of Chicago Press.

Yang, Z., & Purves, D. (2003). A statistical explanation of visual space. *Nature Neuroscience*, 6(6), 632- 640.

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science* 78(2), 238-263.

Zednik, C. (forthcoming). Cognitive Mechanisms. In S. Glennan & P. Illari, *The Routledge Handbook of Mechanisms and Mechanical Philosophy.* London: Routledge.

Zednik, C. & Jäkel, F. (2014). How does Bayesian reverse-engineering work? In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 666-671). Austin, TX: Cognitive Science Society.