CHAPTER

2

Bayesian Psychology and Human Rationality

S. Nichols*, R. Samuels**

*Department of Philosophy, University of Arizona, Tucson, AZ, United States; **Department of Philosophy, The Ohio State University, Columbus, OH, United States

2.1 INTRODUCTION

Human beings make lots of mistakes. It does not take a study to show that when we are drunk, tired, or in the grip of rage, we can believe and do some very silly things. But according to an enormously influential vein of scientific research, one that has dominated the study of human judgment and decision-making for more than four decades, we are error prone in far more fundamental ways. Across a very wide range of judgment and decision-making tasks, people appear to make errors that systematically violate familiar canons of rationality (Baron, 2008; Pohl, in press). This has led many to conclude that the formal theories encoding these canons—the probability calculus and expected utility theory in particular—simply fail to describe human cognition. More generally, the evidence of deep deficiencies in human reasoning has led some philosophers and psychologists to worry that human beings are not, as previously supposed, rational beings at all—that we "lack the correct programs for many important judgmental tasks" and lack "an intellect capable of dealing conceptually with uncertainty" (Slovic et al., 1976, p. 174).

This pessimistic interpretation of the research on human inference is not, of course, without its detractors. One very common response is to criticize, on methodological grounds, the various experiments that are supposed to support such pessimism (Schwarz, 1996; Gigerenzer, 1996). Another is to reject the normative standards typically adopted by proponents of the pessimistic interpretation (Gigerenzer and Gaissmaier, 2011). But perhaps the most influential line of response comes from recent efforts

Rationality. http://dx.doi.org/10.1016/B978-0-12-804600-5.00002-7

to apply Bayesian statistics to cognition. Over the past decade, the development of Bayesian models has become pervasive across the cognitive sciences, including vision science, linguistics, memory research, developmental psychology, and the psychology of reasoning. Although these models vary considerably, one widely shared presumption is that human cognition is, in some quite fundamental sense, well described by Bayesian probability theory. Further, since Bayesian cognitive scientists—in full agreement with proponents of the pessimistic interpretation—view probability theory as a normative theory of rationality, they also contend that human cognition is in some quite fundamental sense rational. As one group of prominent researchers has put it:

> [I]t seems increasingly plausible that human cognition may be explicable in rational probabilistic terms and that, in core domains, human cognition approaches an optimal level of performance. (Chater et al., 2006)

Thus in contrast to the pessimism described earlier, Bayesian cognitive scientists tend to be optimistic when it comes to matters of human rationality.

This paper is part of a larger project in which we chart carefully the implications of Bayesian research in cognitive science for debates over the extent of human rationality. In most general terms, our question is this:

> The Vindication Question: To what extent does recent Bayesian psychological research vindicate the contention that human cognition is rational?

Addressing this question turns on triangulating three kinds of issues: (1) issues about the norms of rationality, (2) issues about the nature of Bayesian cognitive models, and (3) empirical research regarding the fit between these models and actual human performance. In the present paper, we restrict ourselves to clarifying the issue of how Bayesian norms should be construed, and to working through one particular study—due to Tania Lombrozo—which illustrates some of the complexities involved in assessing the implications of Bayesian research for claims about the extent of human rationality. Though the conclusions we reach are by necessity provisional, the position we adopt is neither as pessimistic as some would advocate, nor as optimistic as others. Judgments are more sensitive to evidence than is suggested by the pessimists, but it’s far less than optimal.

2.2 THE STANDARD PICTURE AND THE STANDARD EMPIRICAL CHALLENGE

In order to assess how Bayesian research bears on issues about the extent of human rationality, we need some normative standard against which the quality of human inference can be measured: an account that specifies how one ought to make judgments and decisions. As one might expect, there is considerable debate in both philosophy and the social sciences concerning this issue. Nevertheless, there is widespread consensus among reasoning researchers in general, and Bayesians in particular, that the default standard is what the philosopher Edward Stein has called the standard picture of rationality.[1]

2.2.1 First Pass

According to the standard picture (SP):

> [T]o be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth. If the standard picture of reasoning [rationality] is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with. (Stein, 1996, p. 4)

This characterization of SP is very widely adopted in the literature, often by quoting exactly the passage cited previously. We find a very similar description from psychologists Chase, Hertwig, and Gigerenzer (1998):

> Most researchers of inference share a vision of rationality whose roots trace back to the Enlightenment. This now classical view holds that the laws of human inference are equivalent to the laws of probability and logic (p. 206).[2]

With one significant caveat, which we discuss below, this characterization accurately captures the received view of rationality within the intellectual communities most relevant to our present discussion, though, as we will soon see, it excludes many others. Most importantly, it clearly captures the attitudes of Bayesian cognitive science. In a recent, influential paper, for example, Perfors and coworkers are quite clear that they view logic and probability theory as the normative core of a theory of rationality:

> Bayesian probability theory is not simply a set of ad hoc rules useful for manipulating and evaluating statistical information: it is also the set of unique, consistent rules for conducting plausible inference (Jaynes, 2003). In essence, it is an extension of deductive logic to the case where propositions have degrees of truth or falsity—that is, it is identical to deductive logic if we know all the propositions with 100% certainty. Just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. As Laplace (1816) said, “probability theory is nothing but common sense reduced to calculation.” (Perfors et al., 2011a,b)

In short, not merely do Bayesian cognitive scientists think that probability theory is a powerful descriptive resource, they also maintain that it constitutes a core aspect of a normative theory of rationality. In what follows we assume a conception of SP incorporates Bayesian probability as a part.

Now for the caveat. It is important to notice a lacuna in the previous characterizations of SP. As a matter of fact, the aspects of the rationality debate on which philosophers have tended to focus are those concerned with theoretical reasoning: roughly, reasoning concerned with the making of judgments and revision of belief. The principles of reasoning most relevant to such tasks, and the ones foregrounded by Stein, are those derived from logic and probability theory—hence the reference to "principles of reasoning … based on rules of logic, probability theory and so forth." But the "so forth" covers a class of principles that ought not to be ignored. For there is more to reasoning than theoretical reasoning. In addition, there is practical reasoning, which is concerned not so much with what to believe as with what to do, with the making of decisions. Despite the tendency of philosophers to focus on theoretical reasoning, it is quite clear that those psychologists and behavioral economists interested in human rationality are at least as interested in practical reasoning—with how well we make decisions. And just as there are principles of theoretical reasoning derived from the formal theories, there are also principles of practical reasoning based on formal theories, albeit expected utility theory as opposed to logic or probability theory (von Neumann & Morgenstern, 1944). Indeed much of the most important empirical work on reasoning by Kahneman and Tversky, (1979) among others, has concerned the extent to which human decision-making conforms to the dictates of expected utility theory. In view of this, a more complete characterization of SP ought to make reference to expected utility theory as well as logic and probability theory. This will be important to our discussion in later sections.

2.2.2 Accordance Conditions and the Standard Picture

It is common to note that implicit in SP is a general view about normative standards, sometimes called *deontology* (Stich, 1990; Samuels, Stich, & Bishop, 2002). What deontologists quite generally maintain is that what it is to reason correctly—what is constitutive of good reasoning—is to reason in accord with the appropriate set of rules or principles. The SP adds to this a specification of what the appropriate rules are, viz, ones based on logic, probability theory, etc.

All this is, of course, familiar territory. What is less commonly noted, however, is that this view of rationality has a crucial lacuna: there is no specification of what accordance with the rules requires. The problem is that these accordance conditions can be specified in quite different ways; and different specifications lead to quite different conclusions, both about the plausibility of SP as a normative standard, and about the extent of human rationality. In what follows we consider, and eliminate, two obvious conceptions of accordance before suggesting an alternative, more tenable view, one that we think makes better sense of Bayesian claims regarding human rationality.

2.2.2.1 Accordance as Optimal Performance

Let us start by eliminating a conception of accordance conditions that is obviously too strong. Imagine an agent whose beliefs, inferences, and decisions always conformed to SP. Such an agent would, for example, satisfy the coherence conditions specified by Bayesian probability theory and would always maximize expected utility. The performance of such an agent would accord precisely with that prescribed by SP. It would perform optimally by the lights of SP.

Of course, no one—not even the most ardent Bayesian—claims that humans accord with SP in this way. It is very clear that fatigue, intoxication, distraction, limits of attention and memory, and a host of other factors result in errors. That we make such performance errors is common ground between all parties (for further discussion, see Stein, 1996, Chapter 1, and Stanovich, 1999). That is not to say, of course, that disagreements about the extent of human rationality never concern performance. There are, for example, plenty of disagreements concerning the precise extent to which our inferences and judgments fit the patterns prescribed by SP. But such matters are almost invariably secondary to issues about the extent to which our underlying inferential competences are normatively appropriate (Stein, 1996). Indeed data about performance are typically of central interest only to the extent that they are considered to help assess claims about the nature of this underlying competence.

2.2.2.2 Strong Algorithmic Accordance

Though there are many ways to construe inferential competences,[3] when researchers are interested in whether an inferential process is normatively appropriate they very typically supposes that competences are to be construed as algorithmic level descriptions of psychological processes. This is, for example, what Slovic et al. appear to be assuming in the passage quoted earlier, when they suggest that humans "lack the correct programs for many important judgmental tasks."[4] Suppose this is so, that the relevant level of normative assessment is a Marrian algorithmic level. Then accordance with SP should also be an algorithmic requirement. Further, on such a view it is natural to think that accordance with SP requires some stepwise isomorphism between the mathematics of probability theory and the inferential process under consideration. So, for example, accordance with Bayes rule would require that a cognitive process conform, in a stepwise fashion, to the mathematical operations required to compute Bayes rule.[5]

Though seldom articulated, we suspect the present view, which we call strong algorithmic accordance, is implicit in many discussions of SP, especially among those who reject SP on the basis of familiar tractability considerations (Gigerenzer et al., 1999). As many theorists have noted, executing optimal inferential principles, such as Bayes rule, are extraordinarily, computationally demanding. For example, as Harman notes:

> If one is to be prepared for various possible conditionalizations, then for every proposition P one wants to update, one must already have assigned probabilities to various conjunctions of P together with one or more of the possible evidence propositions and/or their denials... [T]o be prepared for coming to accept or reject any of ten evidence propositions, one would have to record probabilities of over a thousand such conjunctions for each proposition one is interested in updating (Harman, 1986, 25–26)

Thus Bayesian conditionalization is intractable in the technical sense that it is superpolynomial in the size of the input.[6] But more importantly, given the computational demandingness of Bayesian calculations, we can know—even before entering the lab—that people are not doing these calculations. In which case, if the standard picture requires of rational agents that they solve these problems by actually doing the computations, the consequences appear dire for either SP or human rationality. One might accept the characterization of rationality offered by SP but deny that people are rational. Alternatively, one might maintain that SP is mistaken. Indeed, one might do so precisely because SP entails that people are not rational. For as Rysiew notes: "Insofar, then, as we wish to preserve even the possibility that humans are rational… SP seems like a pretty unsatisfactory account of what rationality requires" (Rysiew, 2008, p. 1165).

If the forementioned is correct, then a commitment to both SP and the claim that humans are rational would appear unstable. Specifically it would seem that one cannot insist, as Bayesian cognitive scientists do, that probability theory is normative and that it accurately describes human inferential processes. Yet for all the familiarity of this conclusion, we think it is mistaken. A very different but in our view more sensible reaction is to note that the dilemma turns on an uncharitable reading of SP. The claim that we cannot satisfy the norms of SP turns on assuming a strong algorithmic conception of accordance—that we would need to solve computationally difficult problems, such as belief updating, by actually doing the computations specified by SP. But if this is so, then rather than rejecting SP, we think it is reasonable merely to reject a strong algorithmic conception of SP's accordance conditions. On such a view, there is no need either to reject the rationality of human cognition or to dispense with SP. Rather, what is required is an alternative, more sensible construal of accordance conditions. What might this be?

2.2.2.3 Weak Algorithmic Accordance

In our view, there is a natural answer to this question, which is well motivated by how scientists explicitly handle the task of analyzing large data sets. In brief, scientists routinely confront statistical problems that cannot be solved by analytic methods. To calculate analytically the denominator in Bayes theorem, for example, one needs to sum the joint probabilities of each combination of values from each variable. And in order to do this, the number of joints that need to be calculated increases exponentially as the number of variables increases (eg, if there are 5 variables, each with 4 values, then the number of joints that need to be calculated are 4^5). In problems with many variables, this is intractable, not just for our people, but for our most powerful supercomputers.

In such instances, what do scientists do? What they do not do is throw up their hands and exclaim that no rational means of calculation is available. Instead they develop and deploy various approximation techniques. Over the last 20 years or so, researchers have developed a range of sampling methods that approximate Bayesian inference: for example, Markov Chain Monte Carlo methods such as the Metropolis Hastings algorithm. To get an intuitive sense of how such methods work, imagine there is a box in front of you that contains hundreds of dice of different denominations. Your task is to estimate the average result of a roll of a die randomly taken from the box. The analytic solution would require identifying all of the different dice, their denominations, their biases, and computing the priors for each die type and the likelihoods for each value for each die type. And even for a few dozen dice, this would vastly exceed available computational resources. Here is an alternative, more tractable strategy. Instead of seeking an analytic solution, you could just sample from the box: randomly pull out a die, roll it, record the result, replace, and repeat. This provides you with a sample from the posterior distribution; and if you collect a sufficiently large sample, you can use the average of these values to estimate the true mean. Further, the sample can be used to calculate other features of the probability distribution, such as, the error and standard deviation.

Clearly such an approximation of Bayesian inference is not an analytic solution. In a sense, it does not use Bayesian inference at all. It is not as if these kinds of method use Bayes theorem, for example. Instead, they provide reliable and general methods that enable scientists to bypass the need for analytic solutions. Further, such sampling methods are very typically the best, feasible options available to scientists; and for this reason, they have been used across a broad array of fields, including epidemiology (Hamra, MacLehose, & Richardson, 2013), population genetics (Beaumont, Zhang, & Balding, 2002), and astronomy (Van der Sluys et al., 2008). Further—and this is our main point—no one would seriously deny that it is rational for scientists to use such methods. That is, tacit in scientific practice is the presumption that such methods are rational. Indeed, we suspect that denying this presumption would be viewed by most—ourselves included—as just plain silly.

What does all this have to do with how best to construe SP? If it is rational for scientists to deploy approximation techniques to handle otherwise intractable computational problems, then we maintain it is no less rational for individual cognizers to do so. In other words, we think that, construed algorithmically, accordance with SP should require no more than good approximation methods, at any rate, not when analytic solutions are infeasible. To a first approximation, then, we propose the following construal of accordance:

> Weak Algorithmic Accordance: Where no tractable analytic solution is available, a cognitive process (or system) accords with SP—Bayesian norms, in particular—when it implements a technique that constitutes a good Bayesian approximation method.

This proposal requires some unpacking. First of all, notice that is it less demanding than strong algorithmic accordance in at least two respects. First, it does not require that we possess God-like computational abilities. This is because the runtime properties of good approximation algorithms are, more or less by definition, more feasible than those of optimal solutions. In particular, they are not superpolynomial on the size of the input. Second, though perhaps less obviously, weak algorithmic accordance is less demanding in the sense that it does not require that our inferential competences—absent performance errors—compute the Bayesian optima. Recall, on the strong algorithmic conception, an inferential competence must be isomorphic to the formal principles of SP. But since these principles define the optimal function, it also follows that a rational competence must underwrite optimal computation. In contrast, the requirement that a reasoning process implement a good Bayesian approximation method imposes no such demand, since an approximation algorithm can be very good—indeed even if it systematically deviates from the optima.

So, we have explained two respects in which weak algorithmic accordance yields a less demanding, and more tenable, construal of SP. But we also need to say more about what demands it does impose, specifically what counts as a good Bayesian approximation technique. As one might expect, there is a great deal to be said here. Indeed, there is an enormous literature in theoretical computer science regarding the desiderata on approximation techniques and how best to implement them.[7] Further, there is a very substantial literature on sampling methods, such as Monte Carlo Markov Chain methods and Gibbs filters. But for the moment, we restrict ourselves to four comments.

First, good approximation techniques are developed in such a way as ensure generality. Specifically, approximation methods are almost invariably designed to produce a result across the full range of a problem's instances, where a problem is defined by its optimal solution. In the case of Bayesian sampling methods, the problem is defined by the optimal, that is, Bayesian, means of calculating posterior probabilities. So, good Bayesian approximation techniques reliably approximate the Bayesian optima for a very wide range of cases.

Second, good approximation techniques are very typically capable, subject to resource limitations, of achieving extremely close approximations to the optima. In the case of Bayesian sampling methods, such as the Metropolis–Hastings algorithm, the result asymptotes to the optima as a function of the number of samples that are taken.

Third, and importantly for our purposes, good Bayesian approximation techniques require a sensitivity to large amounts of relevant information. Though they permit tractable computation in part by not considering every available piece of information, the dual demands of generality and close approximation to the Bayesian optima require that such methods sample very widely, and in an unbiased fashion, from the posterior distribution. In this regard, they are quite unlike many of the inferential methods recently popularized by cognitive scientists, such as the fast and frugal heuristics, well known from the work of Gerd Gigerenzer and his collaborators, which we discuss briefly in the next section. For in contrast to Bayesian sampling methods, such heuristics solve judgmental tasks despite ignoring virtually all the available information (Gigerenzer et al., 1999).

Finally, what counts as a good (ie, rational) approximation technique to use can vary across contexts. Imagine two approximation algorithms for the same problem, one is slow but highly accurate, the other is fast but less accurate. In a context where accuracy is highly valued and speed is not, then it is irrational to use the fast approximation algorithm. However, in a context where it is crucial to get an answer quickly, then it can be rational to use the less accurate algorithm. This context sensitivity of what counts as a good approximation technique is naturally accommodated in terms of expected utility. What counts as rational will depend on the utilities associated with solving the task in a particular context. If there is a high utility for speed and lower utility for accuracy, expected utility theory can say that it is rational to use the fast algorithm.

2.3 THE STANDARD CHALLENGE TO HUMAN RATIONALITY

In the previous section, we sought to develop a version of SP that provides guidelines for the assessment of human cognition without being so idealized as to fall afoul of familiar tractability objections. On a weak algorithmic conception of accordance, SP does not guarantee human irrationality. Nonetheless the elaboration of SP does little to help address the most prominent challenge to human rationality. This is because the standard challenge is an empirical one that goes far beyond saying merely that agents are hampered by various processing constraints.

2.3.1 The Challenge (a Reminder)

According to the standard challenge, there is an enormous and growing body of data which suggest that people fail to accord with SP because they systematically ignore critical information in making probabilistic inference. The key tradition here, heuristics and biases (HB), is quite well known, so we will not go into detail here. Rather, we will just present one illustration—but a compelling one—concerning the tendency for people to ignore base rate information. In a classic experiment, Kahneman and Tversky (1973) gave one group of subjects the following scenario:

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

Another group of subjects got the same scenario, but with the base rates reversed; in this condition there were said to be 30 lawyers and 70 engineers. Subjects were then given descriptions, one of which was neutral, another was made to fit with stereotypes of lawyers, and another with the stereotype of engineers. Here is the text for the engineer stereotype:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

Now, subjects are supposed to indicate how likely it is (from 0 to 100) that Jack is an engineer. Kahneman and Tversky found that participants in both conditions gave the same, high probability estimates that Jack is an engineer. The fact that there were 70 engineers in one condition and 30 in the other had no discernible effect on subjects' responses. Moreover, when subjects were given the description that was neutral with respect to the stereotypes, people tended to say that there was a 50% chance that the person was an engineer, once more indicating that they were not using the available base rate information.

Notice: the problem here is not computational tractability. These are simple statistical problems, but people perform poorly at them. And this is just one example taken from a very large set. People show systematically bad performance on both demonstrative and nondemonstrative inference (see Stanovich, 1999, and Pohl, in press, for reviews). Moreover, Kahneman and Tversky have a systematic explanation for these reliable patterns of error: people rely on heuristics that often yield accurate results, but also deviate in systematic ways from rational norms. In the case of the lawyers and engineers problem, for example, people are relying on a representativeness heuristic whereby they estimate probability by thinking about how representative a description is of the category, without integrating the base rate information into their judgment.

2.3.2 A Consensus in the Research on Human Reasoning

The standard challenge to human rationality is very typically developed by drawing on research from the HB tradition. But it is important to note that, despite often intense criticisms, even the most prominent opponents of this tradition are in substantive agreement regarding the extent to which human cognition accords with SP (Samuels et al., 2002). Most notably, the research program associated with Gerd Gigerenzer, which promotes fast and frugal heuristics (FFH), does little to undermine this claim (Gigerenzer et al. 1999). To see why, consider one of the most effective such heuristics: *Take the Best.* Imagine you have to predict which of two cities has a higher rate of homelessness. Further, imagine you have six cues—whether the city has rent control, whether the temperature is above or below median, and so on—and that these cues are ranked in terms of how well they predict rates of homelessness. *Take the Best* says that when predicting which of two cities has the higher homelessness rate, one should initially only look at the best predictor, for example, rent control, and if one city has rent control while the other does not, one should, without considering any further information, judge the city with rent control to have a higher rate of homelessness. Only if the best predictor fails to discriminate—if the two cities both have, or both lack, rent control—should one consider the next best predictor. And only if the second cue fails to discriminate should the third best cue be considered, and so on, down the list of predictors.

Now it turns out that heuristics, such as Take the Best, do quite well on a range of prediction tasks. Indeed, Take the Best often does as well as models that take all of the cues into account (Gigerenzer, Czerlinski, & Martignon, 2002). But for all that, research within the FFH tradition yields much the same conclusion as HB regarding the extent to which human cognition accords with SP. As Michael Bishop notes, according to the FFH tradition, "people can, and often should, use very reliable FFHs that ignore lots of evidence and do not properly integrate the evidence they do consider" (Bishop, 2006, p. 217). In the homelessness case, *Take the Best* counsels us to ignore all the rest of the data once we have found the best cue that discriminates between the cities. This deliberate neglect of data is plainly at odds with SP. More generally, heuristics such as *Take the Best* are much like the heuristics proposed by HB, in that they fail to satisfy traditional epistemic demands on rationality. Here is Rysiew on this point:

> [W]e are capable of certain other, more 'coherence'-oriented forms of cognizing – checking for consistency; deliberately, even ponderously, weighing evidence; reflecting on our belief-forming processes themselves; not to mention, conducting empirical investigations into our own natural belief-forming tendencies so as, perhaps, to ultimately become better thinkers; and so on. And these sorts of more SP-type activities are the sort of thing that many epistemologists have thought to be central to epistemic rationality, and the kind of thing that's required for justified belief and knowledge. (2008, p. 1166)

Sensibly, most epistemologists do not give necessary and sufficient conditions on good reasoning. But as Rysiew's passage suggests, epistemologists often suggest necessary conditions. Internalists, in particular, maintain that good reasoning requires the agent to be attentive to possible inconsistencies among her beliefs and to be sensitive to the available evidence (Cohen, 1986, p. 575).

Despite their myriad disagreements, then, the HB and FFH traditions wholly agree that human cognitive processes very typically fail to satisfy traditional demands on rationality, and as such they agree that we fail to accord with SP. Of course, there are many ways in which philosophers and psychologists have responded to such claims (Samuels et al., 2002). In what follows, however, we want to consider one recent and extremely direct attempt to rebut the challenge. As we will see in the next section, there is a growing body of evidence that suggests that people's inferences do in fact conform to the principles of probability theory.

2.4 RATIONALITY REANIMATED

Recent work in Bayesian cognitive science provides a new possible response to worries about human rationality. In this tradition of work, one identifies a cognitive problem that needs to be solved, and then characterizes the normatively appropriate solution to the problem in terms of standard tools of probability theory, like sampling and model selection. Then one conducts experiments to measure whether human judgment and decision conforms to the normative model. Researchers within this tradition maintain that people draw inferences that conform to Bayesian models across a wide range of cognitive domains, including causal inference (Griffiths & Tenenbaum, 2009), grammar learning (Perfors et al., 2011a,b), and category learning (Kemp, Perfors, & Tenenbaum, 2007). Indeed, several studies have shown that infants make appropriate probabilistic inferences: infants are sensitive to priors (Téglás et al., 2007), are attentive to whether sampling is random or directed (Kushnir, Xu, & Wellman, 2010), and even infer overhypotheses (Dewar & Xu, 2010).

In order to discuss different aspects of how the Bayesian program impacts debate over rationality, we describe in some detail one example from recent research on probability judgments and simplicity in causal explanation. The research we discuss, by Tania Lombrozo and colleagues, draws on Bayesian theory to evaluate human performance. But the research is actually not presented as part of the Bayesian psychology program proper. We focus on it because it is especially apt for considering whether humans exhibit weak algorithmic accordance with SP.

It is a familiar theme in the philosophy of science that simpler hypotheses should be preferred. In the context of probabilistic inference, we can see one reason for this preference. More complex hypotheses risk overfitting the data. The greater flexibility of such hypotheses can mean that they extend to capture aspects of the data that should properly be construed as noise. As a result, the more complex hypothesis might do a poor job of predicting future data. In work on probabilistic inference, this issue is addressed by penalizing more complex hypotheses for their greater flexibility. For instance, there is a Bayesian form of Occam's razor that assigns complex hypotheses a lower prior probability (MacKay, 2003).[8] The data can, of course, overturn the prior probability, with the more complex hypothesis winning out. But the simpler hypothesis is favored at the starting gate. Thus we have two normative claims here: (1) all else equal, people should favor a simpler hypothesis over a more complex one; and (2) people should nonetheless reverse that preference if the data strongly favor the more complex hypothesis.

Extant work indicates that people do favor simpler hypotheses in, for example, category learning (Feldman, 2000; Griffiths, Christian, & Kalish, 2008). We will concentrate, however, on the issue of simplicity in causal explanation. In an elegant line of research, Tania Lombrozo has explored the role of simplicity in people's explanations (diagnoses) of disease when provided information about base rates (2007; Bonawitz & Lombrozo, 2012). Base rates are given by specifying the total size of the population and the *n* later specifying the number of people in the population with each disease. Simplicity is a function of the number of diseases the person might have (1 or 2). Since the proportions are stipulated, it is trivial to do the calculations to see when the simpler explanation is more probable.

The experiments present unfamiliar scenarios. In one experiment, the scenario is set on an alien planet, Zorg, and there are three diseases at issue, Tritchet's syndrome, Morad's disease, and a Humel infection. The symptoms too are unfamiliar (sore minttels, purple spots). The structure of the experiment is that disease 1 causes both symptoms, disease 2 causes one of the symptoms, and disease 3 causes the other symptom. As a result, if an alien presents with both symptoms, D1 will be simpler than the other available explanation, which is that the alien has both D2 and D3. The other factor in the decision is the base rate of the diseases in the population. In Lombrozo's experiment (study 2), the base rate information was explicitly provided to the participants. In all cases, the total population was set at 750. In one condition, each disease is present in 50 individuals; in another condition, D1 is present in 50, D2 is present in 250 individuals, and D3 is present in 220 individuals. There were a total of eight such conditions. In all conditions participants were told about an individual alien who had both symptoms, and they were asked which disease(s) the alien had.

Let us walk through an example. Suppose the incidence of each disease is 50. Since both symptoms are present, the two plausible candidate explanations are that the alien has D1 or both D2 and D3. Given the base rates, the probability that the person has D1 is 50/750, and the probability that she has both D2 and D3 is $50/750 \times 50/750$. This yields a probability ratio of 15 to 1 in favor of the simpler explanation. And, indeed, when participants are in this condition, they overwhelmingly favor the simpler explanation. In another condition, the base rates are 50 for D1, 610 for D2, and 620 for D3. In this condition, given the high base rates for D2 and D3, it is in fact significantly more likely that the alien has both D2 and D3 rather than D1. As a defender of SP would hope, in this condition people are more likely to judge that the individual has D2 and D3 rather than D1.

As noted previously, it is widely accepted that in probabilistic inference, simpler hypotheses should be favored, all else being equal. Earlier work had shown that, at least at some implicit level, people favor simpler hypotheses. Lombrozo's data show that at the explicit level, people also favor simpler explanations. Furthermore, Lombrozo shows that this preference for simpler explanations is moderated by base rate information. If the base rate associated with the more complex explanation is sufficiently high (compared with the simpler explanation), people will favor the more complex explanation. Furthermore, Bonawitz and Lombrozo (2012) find similar results with children, using a task involving colored chips that have different effects on a machine. The red chip causes a toy's light to activate, the green chip activates the toy's fan, and the blue chip activates both. When the child has to determine which chip(s) fell into the machine, they favor the blue chip (simple explanation) unless blue chips are very rare, in which case they favor the explanation that a red and a green chip fell in the machine.

In the foregoing example, people seem to show sensitivity to evidence in ways that would be sanctioned by our weak accordance rendering of SP. Adults and children in these tasks are sensitive to both simplicity and to base rates, as the normative theory says they should be. There is no reason to think that the subjects are throwing away data, as in the FFH cases, nor is there reason to think that the subjects are failing to integrate evidence into their judgments as in the HB cases. Moreover, these patterns of inference seem to be domain general. The tasks are pitched as abstract questions about alien diseases (Lombrozo, 2007) and colored chips (Bonawitz & Lombrozo, 2012).

2.5 RATIONALITY RECHALLENGED

Although a casual glace at the work on Bayesian inference might suggest that people exhibit something close to optimal Bayesian performance, a closer look reveals that this is far from the case. This holds for many of the classic results in the field (Kemp et al., 2007; Schulz et al., 2007; Xu & Tenenbaum, 2007). Since we already have a detailed explanation of Lombrozo's results, we will continue to focus on her work.

People should have a preference for simpler explanations, and, as we saw in the previous section, they do. In addition, people should override that preference if the data sufficiently favor a more complex explanation. Again, as we saw, they do that too. However, we omitted a very important fact about the results. People require far more evidence than they should before they will overturn their preference for the simpler explanation.

In Lombrozo's experiment, when the probability ratio is 15:1 in favor of the simpler explanation, virtually all participants prefer the simpler explanation (that the alien has just the one disease that causes two symptoms). Further, when the ratio is 10:1 in favor of the more complex explanation, the majority of participants favor the more complex explanation. But one key detail that we omitted was this: if people are Bayesian reasoners, we would expect almost everyone in this later situation—when the ratio is 10:1—to favor the more complex explanation. Yet, as a matter of fact, only 60% of participants did. More strikingly, when the ratio is 1:1, so that the objective probability (calculated by base rates and joint probabilities) of the simpler and more complex explanation is exactly the same, 90% of participants favor the simpler explanation (241). And when the ratio is 2:1 in favor of the complex explanation, nearly 70% of adults still favor the simpler explanation. Similar results were found in 5-year-old children (Bonawitz & Lombrozo, 2012). The children preferred the simpler explanation when the ratio was 2:1 in favor of the more complex explanation.[9]

So, despite initial appearances of excellence in human reasoning, performance in Lombrozo's studies is not nearly as close to the Bayesian norm as one might hope. But recall: on our preferred construal, the SP demands only weak algorithmic accordance with Bayesian norms. And to evaluate whether people in Lombrozo's experiments exhibit such accordance, we need to know more about what the algorithmic process might be. One great virtue of the Lombrozo work is that it permits a more precise understanding of the process than that afforded by much work in Bayesian psychology. As mentioned earlier, people have an excessive preference for the simpler explanations. But, as Lombrozo notes, there are two explanations for this divergence from proper Bayesian inference. The first is that participants are underweighting base rates. The second is that people have an overly strong prior bias in favor of simplicity. To place this in context, it is helpful to consider an optimal algorithm. Such an algorithm will describe a particular curve that represents responses as a function of different base rates. Let us call that the Bayesian curve. If people ignore the base rates, then we should not expect their responses to exhibit the same slope as the Bayesian curve. On the other hand, if people have a strong prior bias for simplicity, we would expect that to be manifested as a relatively constant factor that overrates simpler explanations. Of course, people might have both a simplicity bias and a tendency to neglect base rates. However, if people have a strong simplicity bias but do not ignore base rates, then we should expect the data curve to look a lot like the Bayesian curve, knocked up by a constant factor, viz., the prior bias for simplicity. As it turns out, this is precisely what Lombrozo finds. The data curve for her experiment does approximate the Bayesian curve, albeit bumped up by a constant factor (2007, pp.242 and 249).[10]

So, Lombrozo finds that people have an excessive bias for simplicity. Yet we doubt that this bias can be explained as a product of performance errors. Rather, it seems to be a feature of the algorithm itself. This means, of course, that the algorithm fails to provide a very close approximation to the optimal solution; and in that sense, it fails to meet the standards demanded of approximation algorithms in science (such as Metropolis-Hastings), where very close approximations to the optima are to expected.

Still, the process is obviously better than a coin flip. Indeed, the data suggest that the algorithm does reasonably well by the other two conditions we set for weak algorithmic accordance.

First, the algorithm is domain general, it is not dedicated only to solving problems about cheaters or incest. Rather, Lombrozo and colleagues' research indicates that the algorithm is operative in tasks involving diagnosing diseases from symptoms and in tasks involving colored chips activating a toy. This illustrates cross-domain capacity of the algorithm. Moreover, insofar as these studies involve arbitrary factors (eg, colored chips, unfamiliar symptoms of unfamiliar diseases), the algorithm itself looks to be domain general.

Second, and more importantly, the algorithm appears to do quite well by the third condition imposed by weak algorithmic accordance: that algorithms ought to be sensitive to large amounts of relevant information. Recall the algorithms from the HB and FFH traditions. For example, the *Take the Best* heuristic, developed by Gigerenzer and his colleagues is designed to ignore most of the available information. Similarly, the representativeness heuristic described by Kahneman and Tversky is supposed to completely ignore base rate information in the course of generating judgments. The algorithm implicated in Lombrozo's causal explanation tasks is clearly not like these. On the contrary, it is sensitive both to simplicity considerations and to base rate information. Moreover, Lombrozo's evidence indicates that the algorithm does not simply pit simplicity against base rates in a competition model, but actually integrates these two sources of information, leading to a nicely graded response curve.

By the standards of weak algorithmic accordance, then, the algorithm implicated in Lombrozo's task gets a mixed score. It does well by the dimensions of sensitivity and generality, but it does less well by the dimension of approximating the optima. So, how do we answer the question of whether the algorithm counts as SP rational? Without a clear proposal about how closely the algorithm must approximate the optima to count as rational, it is impossible to answer this question. Developing such a proposal is obviously beyond the ambitions of this paper. But it may well be that, at least in certain contexts, algorithms that score as well as the one implicated in Lombozo's task count as rational enough.

Acknowledgments

Both authors contributed equally to this chapter. We are grateful for feedback from members of the audiences at the Institute of European and American Studies Conference on Reason and Rationality, the University of Cincinnatti, and OSU's Center for Cognitive and Brain Sciences, especially Tim Bayne, Ian Gold, Hanti Lin, Tony Chemero, Peter Langland-Hassan, Heidi Maibom, Tom Polger, Angela Potochnik, Andrew Leber, Jay Myung, Zhong-Lin Lu, and Per Sederberg. We would also like to thank Stew Cohen, Juan Comesana, Declan Smithies, Terry Horgan, and Jonathan Weinberg for conversation about these issues. Research for this paper was supported by Office of Naval Research Grant No. #11492159 to SN.

References

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge University Press.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Bishop, M. A. (2006). Fast and frugal heuristics. *Philosophy Compass*, *1*(2), 201–223.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2.6*, 206–214.

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

Cohen, S. (1986). Knowledge and context. *The Journal of Philosophy*, *83*(10), 574–583.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871–1877.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). *Psychological Review*, *103*, 592–596.

Gigerenzer, G., & Todd, P. M. the ABC Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigerenzer, G., Czerlinski, J., & Martignon, L. (2002). How good are fast and frugal heuristics? In T. Gilovich (Ed.), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press: Cambridge.

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68–107.

Hamra, G., MacLehose, R., & Richardson, D. (2013). Markov chain Monte Carlo: An introduction for epidemiologists. *International Journal of Epidemiology*, *42*(2), 627–634.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press, 1986.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

MacKay, D. J. (2003). In *Information theory, inference, and learning algorithms* (Vol. 7). Cambridge: Cambridge University Press.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011a). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011b). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Pohl, R. F. (Ed.). (In Press). *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed.). Hove, UK: Psychology Press.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Rysiew, P. (2008). Rationality disputes—Psychology and epistemology. *Philosophy Compass*, *3*(6), 1153–1176.

Samuels, R., Stich, S., & Bishop, M. (2002). Ending the rationality wars: How to make disputes about human rationality disappear. In R. Elio (Ed.), *Common sense, reasoning, and rationality* (pp. 236–268). Oxford: Oxford University Press.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124.

Schwarz, N. (1996). *Cognition and communication: judgmental biases, research methods and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carol, & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Erlbaum.

Stanovich, K. (1999). Who is Rational? *Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stein, Edward (1996). *Without good reason*. Oxford: Clarendon Press.

Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159.

Van der Sluys, M. V., Röver, C., Stroeer, A., Raymond, V., Mandel, I., Christensen, N., ..., & Vecchio, A. (2008). Gravitational-wave astronomy with inspiral signals of spinning compact-object binaries. *The Astrophysical Journal Letters*, *688*(2), L61.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Williamson, D. P., & Shmoys, D. B. (2011). *The design of approximation algorithms*. Cambridge, UK: Cambridge University Press.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Endnotes

1. Even those who demure from this consensus readily acknowledge that it is the default view.
2. Though it should be noted that Gigerenzer and his collaborators are not themselves advocates of the standard picture (SP).
3. See Stein, 1996.
4. Much the same is true of Stephen Stich's well-known suggestion that the assessment of human rationality within cognitive science is principally an issue about the extent to our psycho-logic is normatively appropriate (Stich, 1990).
5. This is quite similar to what cognitive scientists have in mind when they say that a model and the process being modeled are strongly equivalent (Pylyshyn, 1984). The present suggestion then is roughly equivalent to the claim that accordance with SP norms requires human reasoning processes that are strongly equivalent to normative models of reasoning.
6. Roughly, in the worst case, the number of steps required increases exponentially (or worse) as a function of input size.
7. For an accessible introduction, see Williamson & Shmoys, 2011.
8. Technically, the way this works is a bit subtler. In Bayesian Occam's razor one does not simply assign a lower prior for the more complex hypothesis. Rather, the penalty is naturally represented as occurring in the likelihood term. We can think of the more complex hypothesis as a flexible hypothesis composed of more subhypotheses than the simpler hypothesis. And the total probability for all these subhypotheses cannot be greater than 1. When we calculate the posterior probabilities for the hypotheses, we need to accommodate all of the subhypotheses. In effect, we need to spread out the total probability of 1 across all the different subhypotheses in each hypothesis. Since the flexible hypothesis has more subhypotheses, the probability will be spread out more thinly, effectively leaving each subhypothesis of the flexible hypothesis with relatively lower probability than each subhypothesis in the simpler hypothesis.
9. This simplicity bias diverges even from what would be expected in "probability matching".
10. The adults in the Bonawitz & Lombrozo (2012) studies perform much better and do not show a simplicity bias. It is plausible that this is because in the chips task it is much easier to keep track of the base rates than it is in the aliens task. As a result, adults might not need to rely on simplicity at all to succeed at the task.