The Algorithmic Mind

A Study of Inference in Action

Thomas F. Icard, III

ILLC Dissertation Series DS-2014-02

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Copyright © 2014 by Thomas F. Icard, III

Advisor:	Johan van Benthem
Readers:	Michael E. Bratman
	Noah D. Goodman
	Brian Skyrms
Defense:	November 26, 2013

THE ALGORITHMIC MIND:
A STUDY OF INFERENCE IN ACTION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Thomas F. Icard, III

Stanford Version: March 12, 2014
ILLC Version: August 27, 2014

Abstract

What is the nature of human inference? How does it work, why does it work that way, and how might we like it to work? I advance a framework for answering these questions in tandem, with a rich interplay between normative and descriptive considerations. Specifically, I explore a view of inference based on the idea of *probabilistic sampling*, which is supported by behavioral psychological data and appears to be neurally plausible, and which also engenders a philosophically novel and appealing view of what grounds subjective probability. I then discuss this view in the context of the Bayesian program in cognitive science, proposing a methodology of *boundedly rational analysis*, which particularly exemplifies the normative/descriptive interplay. By taking resource bounds seriously, we can improve and augment the more standard rational analysis strategy. This helps us focus efforts to understand how minds in fact infer, and in turn allows sharpening normative questions about how minds ought to infer. Against this background I explore the phenomenon of *metareasoning*, which arises naturally when discussing resource-limited but representationally sophisticated agents, but which has not been explored in the context of probabilistic approaches to the mind. I propose an analysis of metareasoning in terms of the *value of information*, and explore the consequences of this view for how we should think about inference.

The focus of this dissertation is on implemented (or at least implementable) models of agents, and the role of inference in guiding and driving intelligent action for real, resource-bounded agents. Consequently, a number of the suggestions and claims made are supported by simulation studies.

Preface to the ILLC Version

This version of the dissertation is revised only minimally from the original Stanford version, which is available through the Stanford library. I have added citations of papers that have been published in the intervening months, and corrected two or three typos. Otherwise the documents are identical.

I would like to thank Johan van Benthem for the invitation to publish this in the ILLC Dissertation Series, as well as Marco Vervoort of the ILLC for his help and patience in the process. For general acknowledgments pertaining to the dissertation itself, please see the original Stanford version.

August, 2014
London, England

Contents

List of Tables

List of Figures

Chapter 1

Introduction

William James famously said of the stream of thought:

> Like a bird's life, it seems to be made of an alternation of flights and perchings. [...] The places of flight are filled with thoughts of relations, static or dynamic, that for the most part obtain between the matters contemplated in the periods of comparative rest. (James, 1890, 243)

Some of these "flights between perchings" exemplify relations we might broadly characterize as *inferential*, in so far as the resulting state is, in some sense, a conclusion reached on the basis of the beginning state. This thesis is concerned with such inferential processes, and how they figure into our mental lives.

We will not be concerned with analyzing the concept of inference *per se*, nor with the specific technical interpretations of the term, e.g., in logic or statistics. The ordinary notion is intended here to evoke a large class of mental processes—basic transitions in thought—which can, at least roughly, be construed as leading from some "premise" state to a "conclusion" state. Understanding inference thus broadly, our study will encompass phenomena as diverse as the following:

- Beginning with Helmholtz, it has been common to construe perceptual processes as being inferential, e.g., involving estimation of distance to an object based on visual and haptic cues.

- In categorization, a subject makes an inference about some feature of an object, on the basis of observable features and a history of other observations.
- In social interactions, people are constantly drawing conclusions about others' intentions, beliefs, and other mental states, on the basis of their behavior and background assumptions about what moves people to act in different ways.
- Ordinary predictions about future events, e.g., whether a given candidate will win an election, are paradigm examples of uncertain inference.
- Some inferences follow clear logical patterns, such as *modus ponens* or *modus tollendo ponens*. To take an example from van Benthem (2007), of disjunctive inference: a waiter bringing three dishes to a table of three, after finding out who has the first two dishes, need not inquire who has the third.
- Being presented with a decision problem, and concluding that some course of action is the thing to do, we can sometimes understand as inferential.

All of these involve, in some way or another, *going beyond* the information explicitly present in the initial state or observation. With this liberal interpretation, encompassing these otherwise disparate phenomena, understanding the nature of inference can be seen as one of the primary goals of cognitive science.

1.1 Questions about Inference

I find it useful to distinguish three types of questions one can ask about inference: *what?*-, *how?*-, and *why?*-questions. I believe these ought to be answered in tandem—this will be one of the themes of the dissertation—and that a thorough understanding of inference requires asking and answering all three kinds of questions.

1.1.1 What?

What?-questions are the most commonly discussed in philosophical analyses of inference, particularly of *rational* inference. One is interested in relations that hold between expressions, or perhaps more abstract *contents*; and in understanding *what follows from what*, independent of what goes on in people's heads when they infer. Indeed, studying the abstract rules of inference in this way has been taken by many to be the main subject matter of *logic*, Frege being one of the classic modern sources of this idea. Sometime between the 1880s and 1890s, Frege wrote,

> To make a judgment because we are cognizant of other truths as providing a justification for it is known as *inferring*. There are laws governing this kind of justification, and to set up these laws of valid inference is the goal of logic. (Frege 1979, 3; first sentence quoted in Boghossian 2014)

In other words, the goal of logic is to determine what follows from what, in such a way that we can tell when proceeding from some premises to a conclusion would be justified, i.e., when such an inference would be valid. Logical validity is of course not the only important justificatory relation. Even when a conclusion is not an inevitable consequence of some body of information, the information may nonetheless support the conclusion to some degree. Various *probabilistic* or *statistical* relations have also been proposed for this kind of inference, including notions of statistical support, measures of partial confirmation, and others (see Eells 2005 for a survey). Some have also argued for recognition of a distinct kind of *abductive* inference, or *inference to the best explanation* (what C. S. Pierce called "retroduction"), in addition.

The many tomes that have been written on these topics—in philosophy, logic, and statistics—attest to their importance, and there certainly remain many unsettled issues. However, I will have little new to say directly about them in this thesis.

1.1.2 How?

If the *what?*-question is about how the contents of the premise and conclusion states relate, the *how?*-question is about what kinds of concrete—even physical, or at least physically inspired virtual—mechanisms could undergo such inferential transitions. There is an obvious practical component to the idea of building artifacts capable of reasoning, for constructing anything from calculators to dialog bots. To that extent, it is useful to understand how inferences can be carried out in actual matter.

The *how?*-question is equally important for understanding the nature of human inference. As Pierce (1887) noted, already in a discussion of the early "logical machines" capable of performing syllogistic inferences, "the study of [these machines] can at any rate not fail to throw needed light on the nature of the reasoning process" (71). If one is able to support a proposal about how inference works by building an artifact that embodies it, this proves that the proposal is at least feasible, in the sense that the computations required can actually be carried out by *some* physically instantiated system. Moreover, possession of such an artifact allows testing the proposal under different conditions and in different contexts.

In the age of digital computers, these advantages extend to *models of* reasoning artifacts, whereby behavior can be *simulated* on a computer. This is possible even when the physical process (e.g., in the brain) is assumed to work quite differently from the way a digital computer works, in detail. In psychology, research programs that focus on mechanisms underlying inference have often relied on such techniques (e.g., work on neural network models, Rumelhart et al. 1986, *inter alia*).

This thesis does aim to make contributions to how we understand the *how?* of at least some types of inference, as I will explain shortly.

1.1.3 Why?

Much ink has been spilt over the *what?*- and *how?*-questions of inference, including where they interact, viz. what kinds of artifacts can carry out which kinds of inferences. There has been much less discussion of *why?*-questions. This is surprising. Even if we had a perfect understanding of how a physical mechanism (even a brain) could carry out different types of inferences, and even if we had a perfect catalog of all the different notions of "what follows from what" and when each is appropriate, this would still tell us precious little about what an agent in a particular situation *ought* to infer. Likewise, it would not necessarily allow us to *predict* what an agent *would* infer in any given situation.

The quotation above from Frege states that, in inferring, we make a judgment "because we are cognizant of other truths as providing a justification for it". That is, the reason I infer B from $A_1, \ldots, A_n$ is that I see that B follows from $A_1, \ldots, A_n$.[1] This, of course, cannot be the whole story. From anything many things follow. The *why?*-question asks us to consider why an agent would make one "flight between perchings" rather than another. What (rationally) *directs* inferential processes?

Although there has been little philosophical or theoretical progress on this *why?*-question, it has not gone unnoticed. In his seminal 1950 *Mind* paper, Turing articulated the problem poignantly, in a passage I quote in full:

> At each stage when one is using a logical system, there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a footling reasoner, not the difference between a sound and a fallacious one. Propositions leading to imperatives of this kind might be "When Socrates is mentioned, use the syllogism in Barbara" or "If one method has been proved to be quicker than another, do not use the slower method". Some of these may be "given by authority", but others may be produced by the machine itself, e.g., by scientific induction. (Turing, 1950, 458)

The *what?*-question may distinguish a sound from a fallacious reasoner, but we need to answer the *why?*-question to distinguish a brilliant from a footling reasoner. As Turing also makes clear, the *why?*-question is directly relevant to the *how?*-question, and specifically the question of how to build an artifact that makes intelligent inferences. As he says, a partial solution to this problem may be given "by instruction" to the agent, while other aspects must be learned by the machine itself, foreshadowing the idea that an agent needs to be able to *learn to reason.* An agent does not merely need to know a logical calculus or how to make statistical hypotheses; somehow, it needs to know how to direct its reasoning in an intelligent way.

Answering the *why?*-question may seem to be an intractable task. It bears some resemblance to (perhaps even generalizes) a version of what has been called *the frame*

[1] In the recent philosophical literature, both Broome (2013) and Boghossian (2014) have endorsed this as a necessary condition for inference. Broome adds the condition that the agent "care about the conclusion" (Ch. 13), which seems on the right track to addressing the *why?*-question.

problem in philosophy of mind and artificial intelligence (see Dennett 1981a; Fodor 1983; Pylyshyn 1987): deciding what is and what is not relevant to a given reasoning problem. If we could answer the *why?*-question, would we not thereby solve the general problem of how to reason effectively and efficiently?

One could raise an analogous objection to the idea that we ever could have made progress on the *what?*-question. After all, in many domains the question of what follows from what is provably undecidable. This does not stop us from making progress in our understanding, e.g., of first-order entailment, or arithmetical consequence. We can gain a better grasp of the concepts and notions involved without being able to determine in any given case whether some consequence in fact follows. Likewise, we would like to understand the *why?*-question—its "logic", so to speak—even if we cannot easily determine in any given case what the rational inference step would be.

1.2 Inference in Action, Inference as Action

The strategy I propose for making progress on the *why?*-question is to understand inference as a special case of more general rational activity. In short, the proposal is that it makes sense to undergo one inferential transition rather than another to the extent that the first is *more conducive to the goals of the agent.* In other words, inference, in the general case, is to be viewed instrumentally, as merely a means to an end, the end being whatever goal a person may be pursuing. This, too, is an old idea, again taking a cue from the pragmatists. As Stich has put it:[2]

> For the pragmatist, cognitive mechanisms or processes are to be viewed as tools or policies and evaluated in much the same way that we evaluate other tools and policies. One system of cognitive mechanisms is preferable to another if, in using it, we are more likely to achieve those things that we intrinsically value. (Stich, 1990, 24)

[2] See also the related discussion in Godfrey-Smith (1998) of Dewey's idea that minds are "for dealing with complex environments"; though he additionally argues Dewey held that certain mental states could be ends in themselves.

This suggests a general way of assessing cognitive architectures and mechanisms. Slightly more concretely, I propose that we understand mental processes in the context of an *underlying decision problem*: something the agent is assumed to be occupied with, working toward, or otherwise aimed at. A cognitive mechanism will be more or less rational to the extent that it contributes to the agent's maximizing this end.

For the specific case of *inferential* processes—which, again, are distinguished roughly by their leading from one state to another, the relation between them being described in terms of the *what?*-question—this already marks some progress on the *why?*-question. We find ourselves making certain inferential transitions rather than others, because they help us achieve some goal, or make progress toward some end.

Consider van Benthem's (2007) waiter example, mentioned above. There is an obvious goal in this case, of putting the right dish in front of the right person. Upon receiving information about where the first two dishes are to go, the waiter concludes that the remaining person ordered the remaining dish. Whereas there is no limit to what could reasonably be inferred based on the information received, at least as far as the *what?*-question is concerned, the reason the waiter makes this specific inference is that it enables taking the correct action to achieve the underlying goal.

As van Benthem stresses, the role of *questions* in this scenario is central. Appropriate action cannot be taken without specific information about the situation, viz. who ordered what dish. Since the waiter currently lacks this information, the instrumentally rational thing to do is *pose a series of questions* to find out. These are of course verbalized questions to other agents. Taking this idea further, I submit that the very act of inferring involves *posing a question to oneself.* In the scenario, the conclusion that the third person has the third dish is reached as an *answer* to the question, "Who has the third dish?", which is the remaining piece of information needed to take appropriate action for meeting the waiter's goals.

Viewing inference as involving a question to oneself has two attractive consequences. The first is that it captures the intuition that, in inferring, we do not know where we are going until we get there. James described the phenomenology particularly poetically:

> If they are but flights to conclusions, stopping to look at them before the

> conclusion is reached is really annihilating them. [. . .] The rush of the thought is so headlong that it almost always brings us up at the conclusion before we can rest it. (James, 1890, 243)

As I will argue further in Chapter 4, the whole point of inference is to bring to light answers to questions, which are relevant to the current problem being solved, and which the agent already has the resources to provide. In many cases, as I will discuss at length in Chapter 2, inferential processes can be *essentially nondeterministic*, meaning that even an outside observer could not know beforehand what conclusion will be reached. *Pace* the quotation from Frege above, inference as a mental process does not involve considering whether a conclusion follows, though such activity may certainly succeed the inference;[3] once the conclusion is reached, the inference has been made. The process is driven by the question posed.

The second attractive feature of this view, I claim, is that it allows us to locate the answer to the *why?*-question in the analysis of what makes for a good question. Faced with a decision problem, some strategies for approaching the problem will be better than others. In the waiter scenario, asking oneself who has the third dish is obviously most helpful for the task of putting the plate at the right place. In other cases, it may be less obvious what the best strategy of (self-)inquiry may be. Fortunately, there is already a large body of literature on the *value of information*, or the *value of experiments*, which studies information gathering in the context of a decision problem (e.g., Raiffa and Schlaifer 1961). It has been suggested that similar ideas can also be used to analyze questions in natural language (van Rooij, 2004). In Chapter 4, I will argue that we can appropriate these same tools by *internalizing* them, allowing us to analyze strategies of inquiry and deliberation, which in turn can be thought of as *patterns of self-questioning* leading to individual inferences. We find ourselves making more or less rational inferences—that is, making the “right” inferences in the “right” circumstances—because we have effectively solved the value of information problem for harvesting our own minds.

[3] There is not necessarily a substantive disagreement here. I am understanding inference as a basic mental process, rather than, say, a drawn out activity of considering whether some statement ought to be concluded from some other statements.

All of this raises a question about the extent to which inference on this picture is genuinely *active*. This is an enormous and fascinating topic (see, e.g., Broome 2013 for a lively current discussion), but at least a few words are in order.

Looking at the examples of inference mentioned above, some are clearly not of a sort that *the agent* actively carries out. Rather, they are performed by some subpart of the agent, e.g., in the visual system, and totally beyond the agent's control. Others appear much more agential. Inferring that some course of action is the thing to do is paradigmatically agential, for instance. In some cases, one may be able to guide the inferential process, for example, by deliberately posing oneself questions. As with other intentional action, the process need not be guided all the way to termination. Just as a tennis serve is under the agent's control only up until the follow-through, some inferences may be under the agent's control up until the question is posed. From there, the mind simply comes back with a proposed answer, which is experienced as the outcome of the inference.

To take a well known example from ordinary conversation (Grice, 1975), suppose I am on the side of the road with an immobile car and the following exchange occurs:

Q: I am out of petrol.

A: There is a garage round the corner

Suppose further that, at first, I do not understand why the passerby would have offered this response. I may then have to ask myself, to prod my own mind actively, about what I ought to conclude from this short exchange. In the normal case, the answer will come right away, that the person meant to tell me I can find petrol there. In other cases, solving the problem—in this case, interpreting the utterance—will require yet further prodding. And in other cases, perhaps most other cases, this will all happen automatically, even below the radar of conscious awareness.

Thus, we have a spectrum of cases, ranging from inferences that occur wholly "subpersonally", to cases that involve intentional action on the part of the inferrer. Nonetheless, all of these cases share certain features that are at least "action-like". Following Frankfurt (1978), we can distinguish activity that counts as genuine action (intentional or not) from the broader category of *purposive* activity.[4] By conceiving of inference instrumentally, we are assuming that inference is always purposive. It is always to be thought of as directed toward some end. Moreover, purposive activity has enough in common with genuine action that we can use some of the same tools to understand both. Inference costs time and energy; it comes with opportunity cost, in the sense that inferring one thing precludes inferring something else at the same time; and perhaps most importantly, different inferences may have differential impact on how a person is able to solve the underlying problem at that moment.

To conclude this section: as an approach to the *why?*-question, we are viewing the function of inference as primarily to situate an agent so that it may take more intelligent actions. At the same time, we are considering inference itself as a kind of (at least quasi-)action, so that we can analyze the relative rationality of alternative inferential strategies or mechanisms.

1.3 Bounds, Algorithms, and Rational Analysis

The very idea of inference seems to presuppose that we are dealing with *bounded* agents: agents who have not yet "made it" to some conclusion, or who do not yet know what they ought to do in a given situation. For real, concrete agents, reasoning and inference always require *doing something.* With reasoning, as with any other human activity, how we ought to carry out this activity depends on what we are able to do: "ought implies can", as the slogan goes. That is, to say what a good reasoning strategy, or rational inference, would be requires knowing what kinds of beings we are, and what kinds of limitations bound us.

This is of course an old point, made famous especially by Herbert Simon (1955; 1956). Simon was also adamant that a good way to make progress in understanding human reasoning and inference is to explore concrete *algorithms*, whose behavior is

[4] "Behavior is purposive when its course is subject to adjustments which compensate for the effects of forces which would otherwise interfere with the course of the behavior, and when the occurrence of these adjustments is not explainable by what explains the state of affairs that elicits them." (Frankfurt, 1978, 160) We of course allow, as does Frankfurt, that this adjustment, or guidance, is achieved by subpersonal mechanisms. On Frankfurt's view, it counts as *bona fide* action when such guidance is achieved by the agent.

supposed to mirror the processes of human minds. As he put in, "The moment of truth is a running program" (Simon 1995, 96, cf. also Newell and Simon 1976). Implementing these algorithms requires us to be honest about boundedness, and ensure that we do not smuggle in any unrealistic assumptions about what we think agents ought to be capable of.

In this dissertation, I share Simon's focus on boundedness and concrete algorithms. However, I believe that considerable progress can be made by analyzing *why?*-questions in addition. Apart from being of interest in their own right, I believe *why?*-questions can help us answer *how?*-questions. I submit that we can sharpen our understanding of the *how?*-questions—how reasoning minds in fact work—by asking, e.g., why a reasoner subject to resource bounds would employ those resources in one way rather than another to achieve an aim, or even what the optimal use of resources would be.

This idea, of hypothesizing an *is* from an assumed *ought*, is also not new. A number of philosophers have argued, in one way or another, that a good way of predicting what an intelligent agent will do is to figure out what it would be *rational* for them to do (Lewis, 1974; Davidson, 1975; Dennett, 1981b). More recently, the psychologist John Anderson has proposed a general methodology of *rational analysis*, premised on the assumption that considerations of how a given cognitive mechanism *ought* to work will often lead to good models of how the mechanism does *in fact* work (Anderson, 1990). Most of these proposals, while acknowledging boundedness, do not take bounds seriously enough, however. Particularly in work by Anderson and many following him, models of cognitive processes are usually couched at what Marr (1982) famously called the "computational level", which in my terminology addresses only the *what?*-question. For inferential processes, this means we only look at abstract informational support relations and notions of "what follows from what"; most common is to understand inference in terms of *Bayes Rule*, so that the (rational) conclusion of an inference is the response with highest posterior probability, a claim typically made without consideration of the costs involved in computing the relevant probabilities. While I do believe there is considerable merit to this approach (see Chapter 3), I also believe we can improve upon it by taking bounds more seriously.

Thus, we have Simon and followers on the one side, focusing almost exclusively on *how?*-questions; and we have Anderson and followers on the other side, focusing almost exclusively on *what?*-questions. We should think about both of these kinds of questions, but not to the exclusion of *why?*-questions. Overemphasis on *how?* blinds us to what a given mechanism is doing at a higher level of description, and potentially misses out on insights to be gained from cases where we can reasonably infer *is* from *ought*. Overemphasis on *what?* leads to an inflated and potentially unrealistic idea of what it is rational for a person or process to do, and thus risks being irrelevant to actual human psychology. My understanding of the *why?*-question is precisely about *boundedly rational* inference, and thus strikes a middle ground between these two extreme stances. Considering *why?*-questions in fact already requires us to attend to both *what?*- and *how?*-questions, since boundedly rational inference for an agent can depend both on "what follows from what" and how the agent is in fact constituted.

In order to understand how a given cognitive (inferential) mechanism does (or can) work, it is evidently helpful to know how it *ought* to work. But we can gain a better understanding of how the mechanism ought to work to the extent that we know something about how it does (or can) work. In order to make this potentially vicious circle virtuous, we must find a reasonable starting point.

1.4 Why Bayesian Psychology?

Luckily, we do not have to start from scratch. We can build on everything we already know about how our minds work and how they can work. The problem of inference in particular has been the subject of much investigation, ranging from examination of the *what?*-question, measuring people's performance against purportedly normative measures, to mechanistic theories addressing the *how?*-question.

For a number of reasons, my focus in this dissertation will be on what is known as the *Bayesian program* in psychology and cognitive science (see, e.g., Oaksford and Chater 2007, Ch. 2, 4 for an overview). Roughly speaking, this program analyzes human inference in terms of *Bayesian inference*, and thus, as noted above, has typically been concerned with what I call the *what?*-question. Following Anderson's rational analysis strategy, a typical analysis begins with the "problem being solved", formulated in terms of a prior probability space and some "data" which transforms ("conditions") the space, so that the resulting view—the "conclusion" of the inference—is represented by some aspect of the conditioned space, e.g., the proposition with highest posterior probability.

This program has experienced an explosion of activity in the past 15 years, and the range of cognitive phenomena that have received Bayesian analyses is arguably unprecedented in scope. In addition to nearly all of the "classical" topics of psychological research—from low-level vision and psychophysics (Knill and Pouget, 2004) to categorization, memory (Anderson, 1990), and reasoning (Oaksford and Chater, 2007)—the program has also been extended to a wide range of "higher-level" cognitive phenomena, which had otherwise been recalcitrant to mathematical or computational theorizing, including sophisticated causal inferences (Gopnik et al., 2004), intuitive physics (Sanborn et al., 2013), and even aspects of social cognition and language understanding (Frank and Goodman, 2012). Bayesian models provide a convenient *lingua franca* for inference, and cognition more broadly, by taking a "top-down" approach and demanding a careful, mathematically explicit specification of the problem space and representation language (Tenenbaum et al., 2011).

This latter feature of Bayesian models makes them particularly attractive from the point of view of understanding inference. The "top-down, problem-first" methodology has led to hints of a rapprochement between statistical and probabilistic models of cognition on the one hand, and structured, logical approaches to inference and reasoning on the other (Goodman et al., 2008; Tenenbaum et al., 2011). It is precisely the marriage of these two perspectives that has led to much of the recent progress on highly structured high-level cognition. These developments ought to be welcomed by the philosophical logician interested in closer ties with precise, quantitative, testable cognitive models (see Icard 2013 for further discussion of this point).

While the Bayesian program, broadly conceived, makes for one convenient starting point in theorizing about human inference, as explained in the previous section, the "standard" or "orthodox" program focuses on the *what?*-question almost to the exclusion of the *how?*- and *why?*-questions. Similar worries have been expressed in a recent wave of critical reviews of the Bayesian program by both philosophers and psychologists (see Chapter 3).

One particularly promising proposal for bridging Bayesian models with more mechanistic considerations—that is, connecting *what?* and *how?*—has come from a line of work related to the idea of probabilistic *sampling*. Developed originally in engineering and statistics, sampling methods provide convenient ways of *approximating* difficult probabilistic calculations, such as computing conditional distributions. A number of psychologists have recently proposed viewing inference as a kind of sampling process, and several subtle psychological findings have been explained by appeal to specific sampling algorithms (see Chapter 2). Moreover, sampling algorithms are designed precisely so as to be implementable by a concrete computing device, and in fact there are surprising connections between some well studied sampling algorithms and proposals for how brains actually compute (§2.8, Appendix A.1). All of this makes one hopeful that we may be catching a glimpse of one place where "top down" and "bottom up" approaches to cognition will finally meet.

Speculative as this is given our current state of knowledge, there are two reasons I find it useful to take the Sampling Hypothesis as a tentative working conjecture, apart from its connection to the Bayesian program and the associated *what?*-questions. The first is that it provides a precise, and at least plausibly realistic, model for the actual *process* of inference. It allows us to view inference as genuinely *algorithmic*, and in particular it allows us to substantiate the idea of inference as *posing a question to oneself*. Drawing samples can naturally be viewed as harvesting one's own mind for implicit information. Second, and partly because of this algorithmic emphasis, it permits us to study the *why?*-question in a concrete way. The Sampling Hypothesis raises distinctive questions about rational inference, which make vivid what it would mean for one inferential process to be preferable to another, thus illustrating a concrete example of how we might explore the *why?* of inference.

All three of the main chapters include what I see as important contributions, independent of the Bayesian program, in ways I will explain. At the same time, many of these contributions, I believe, also shed light on the Bayesian program itself, and can be used to defend it, and potentially even improve and advance it.

1.5 Outline of the Dissertation

The project described in this introduction is obviously ambitious, and will require no small amount of work to defend and validate in detail. In this dissertation, I focus on just a small part of the project. My main aim is to make some progress on understanding what I have called the *why?* of inference: why we find ourselves making certain inferential transitions rather than others, and what the rational basis of this could be. This requires saying something about how inference might in fact work, and how a (boundedly) rational agent would use its resources to ensure its inferential strategies conduce to its goals.

In Chapter 2, I explicate the Sampling Hypothesis as I see it, drawing on a diverse body of empirical literature from neuroscience and cognitive psychology. The hypothesis, I claim, provides a novel way of thinking about *subjective* or *personal probability*, a notion that plays a central role in many accounts of cognition and intelligent (inter)action, across many fields. I believe the resulting viewpoint ought to shift the role probabilistic beliefs play in our understanding of (at least instrumental) rationality, underscoring the importance of resource limitations.

In Chapter 3, I explore what it means for an agent (or architecture) to be boundedly rationality, and use this study to defend (and hopefully augment) the Bayesian program in psychology. Specifically, I propose that Anderson's rational analysis methodology for psychology be extended to *boundedly rational* analysis, maintaining many elements of the original strategy, but also incorporating what we know about the concrete limitations and bounds that constrain actual human agents. Drawing on a recent study by Sanborn et al. (2010) on categorization, I illustrate how this methodology would work in a concrete case. Specifically, I show through simulation study that Sanborn et al.'s *particle filter* model, a specific example of a sampling algorithm, is more boundedly rational than several previously studied alternatives. The boundedly rational analysis methodology, I argue, goes some way toward addressing a number of criticisms of the Bayesian approach to cognition.

Considerations of bounded rationality, and the idea that different inferential, or more generally deliberative, strategies may be differentially beneficial, raises the obvious question of how an agent could ever hope to use the "right" strategies in the "right" circumstances. Some of this may be inborn, whether through evolution or general learning (cf. the quotation above from Turing 1950), but it is also possible—and may sometimes be necessary—for the agent to solve this problem *online.* That is, it may be boundedly rational for agents to spend time reasoning about their own reasoning, before continuing on with one or another strategy. As I explain at the end of Chapter 3, this raises some potential puzzles for the boundedly rational analysis methodology, since it would seem we have to consider a very large space of possible agent architectures. I show through computer simulation that there are simple cases where, among sampling agents, those who spend time reasoning about how many samples to take outperform agents who draw fixed numbers of samples. This poses a puzzle for the boundedly rational analysis methodology, but it also highlights the importance of metareasoning, whose study I believe can help solve the puzzle.

Thus, finally, in Chapter 4, I take up the problem of metareasoning as a topic in and of itself. The very idea of metareasoning has been seen as problematic by many, and there has been some worry that any attempt to capture it formally would result in paradox or regress. Taking a cue from I. J. Good and several following him, I propose we understand metareasoning (and in fact inference and reasoning more generally) in terms of *value of information*, turned inward toward our own minds. Reasoning about sampling provides one simple and illustrative example of the proposal. Importantly, these methods are applicable at a very low level of automatic, subpersonal processes, as well as at a high level of explicit, active deliberation.

To be sure, one reason to care about the *why?* of inference is to understand how we can become better reasoners. In so far as this may require us to reason about our own reasoning—or to deliberate about how best to deliberate—it is worthwhile to understand the logic of metareasoning. Thus, also in Chapter 4, I explore how we can make sense of boundedly rational metareasoning, in a way that does not lead to regress. I also distinguish different ways of using the value of information framework.

While later chapters draw upon results and ideas developed in earlier chapters, each chapter ought to stand alone with only a minimum of background. Some of the empirical results discussed and interpreted, I believe, will be of philosophical interest beyond the project I have described here, and I have tried to make my presentation and explanation of these results clear and accessible. Conversely, I hope some of the observations made and questions raised concerning psychological results and programs, and philosophical perspectives thereon, will also be of some independent interest. Yet, my primary goal is to gain some purchase on understanding the *why?* of inference—why we would make certain "flights between perchings" rather than others—and in that, I aim to follow James' advice, in another context, to turn "towards concreteness and adequacy, towards facts, towards action" (James, 1907).

Chapter 2

Subjective Probability as Sampling Propensity

2.1 Introduction

Ramsey held that “a degree of belief is a causal property of it” (Ramsey, 1931), and in particular that subjective probability could be understood as a kind of propensity. The causal property Ramsey had in mind was something, “which we can express vaguely as the extent to which we are prepared to act on it” (71). Subjective probabilities are conceived as propensities to act, and as such become amenable to measurement, e.g., *via* a representation theorem using observed preferences between gambles. Later, de Finetti was even more adamant about tying subjective probability to some measurable phenomenon, sometimes even suggesting we define subjective probability operationally (de Finetti, 1974):

> In order to give an effective meaning to a notion, and not just an appearance of such in a metaphysical-verbalistic sense, an operational definition is required. By this we mean a definition based on a criterion which allows us to measure it. (76)

While the desire to ground subjective probability in something more objective is admirable, the very close link between probabilities and behavior that comes out of the betting interpretation is widely seen as problematic. The problems are familiar, and I will not rehearse them here (see Eriksson and Hájek 2007 for a recent discussion). Many can be traced to a basic issue that subjects' actions are not directly caused by their informational attitudes, but only in concert with motivational attitudes and a decision rule, which itself may depend on other attitudes, e.g., toward risk. Thus, there is an epistemological problem about how we might reasonably infer an agent's probabilities from that agent's behavior, if we have to solve for all of these variables simultaneously. It may seem a leap of faith to assume an agent's mental state can even be factored cleanly into separate informational and motivational attitudes.

One response to some of these worries is that taken by Lewis (1974) and Davidson (1975), who suggest that we should maintain the basic picture, but admit that it plays a largely theoretical role in how we would ideally rationalize an agent's behavior. On this response, we make no claim that some concrete representation of probability has an actual causal role in producing behavior. A more radical response would be to give up subjective probability altogether in our theorizing about subjects' mental states. Both responses, in effect, entail a rejection of Ramsey's dictum, and a kind of antirealism (or at least non-realism) about subjective probability. (See Zynda 2000 and Eriksson and Hájek 2007 for variations on, and refinements of, these positions.)

I believe there is something right about Ramsey's insistence that subjective probability should play a causal role in producing behavior, and that it is, at least in principle, the sort of thing we can measure. However, much in line with the intellectual atmosphere of the time, Ramsey, de Finetti, and many following them were reluctant to posit, or even mention the possibility of, internal mental mechanisms that could play this causal role. This limited them to relying only on certain kinds of behavior. We are now in a position to look beyond this basic behavioral data to other sources of evidence, including potentially richer behavioral sources, and especially evidence stemming from computational modeling and even neuroscience. Indeed, we have numerous means of testing the hypothesis that something like subjective probability is actually represented and used by the mind, and exploring the precise ways that this may be accomplished. Empirical psychologists and cognitive scientists have been doing just that over the past half century, with substantial progress in the last decade. We can follow Ramsey and de Finetti in assuming subjective probability should be tied to something concrete and empirically measurable—in this sense, we can be *realists* about subjective probability—but by grounding it to concrete mechanisms in the mind we have rich additional sources of evidence at our disposal, and the ability to state concrete, testable hypotheses.

My aim in this chapter is to present a general hypothesis about one important way subjective probability seems to be represented. In short, the idea is that the probability of some event or proposition is represented by the propensity of an internal generative model to return that event as output; in other words, roughly by the proportion of times the model would be expected to *sample* that event from among the possible outputs of the model. That is, we replace Ramsey's *propensity to act* with an internal *sampling propensity*. This *Sampling Hypothesis* has received some impressive empirical support over the last several years, and as I shall argue, it has some attractive philosophical consequences. While there are various ways of filling in details of the hypothesis, and of assessing the extent of its applicability to human cognition and reasoning, in this chapter I will mostly remain at a relatively high level of description. My view of the Sampling Hypothesis is compatible with many theoretical programs within philosophy and cognitive science, some of which may otherwise stand in opposition. At this level, I believe the evidence so far makes for a powerful case that we should take the view seriously. If it turns out to be on the right track, it has significant consequences for how we conceive of subjective probability and its role in thought. The resulting high-level picture answers to several longstanding criticisms and reservations concerning the idea that probabilities are represented in the mind, including worries about the complexity and intractability of probabilistic computation, and about the very idea that our minds could harbor precise numerical values representing probabilities. At the same time, the resulting view looks somewhat different from the picture of subjective probability that we inherit from the tradition following Ramsey, de Finetti, Savage, etc., as I will try to make clear.

The cognitive and neuroscientific literature on this topic is complex and growing rapidly. This chapter will not contain a comprehensive survey. Instead, I want to give a general sense of the empirical support for the hypothesis, and especially to explore some of the consequences of the overall view for how we should think about subjective probability. One of the most intriguing features of sampling is that it seems to be a general strategy the mind employs across multiple levels, from relatively automatic, “subpersonal” processes in vision, syntactic parsing, and category learning, all the way to high-level judgments and predictions. In what follows I will move from one domain to another, sometimes glossing over otherwise important differences.

After some preliminary remarks on the intended subject matter (§2.2) and rehearsal of some well-known challenges to subjective probability (§2.3), in §2.4 I shall present the Sampling Hypothesis as I see it, exhibiting the basic idea with an example inspired by Rumelhart et al. (1986). §2.5 will consider an illustrative formalization of the example using a so called Boltzmann Machine, which implements a well known sampling algorithm from the statistics literature, as will be demonstrated in technical Appendix A.1. Following this will be some clarifications of the view (§2.6-§2.7), and in §2.8-§2.10 I will present an overview of empirical evidence supporting the hypothesis, ranging from vision to ordinary prediction. A second technical Appendix A.2 contains a discussion of the relation between the softmax choice rule and a sample-based choice rule, which comes up in the context of probability matching (§2.10). After a short discussion of the role of *value* or *utility* on this picture (§2.11), I will offer some remarks on the problem of measurement from the point of view of the Sampling Hypothesis (§2.12). Finally, in §2.13 will be some concluding remarks.

2.2 ‘Subjective’ versus ‘Psychological Probability’

To avoid confusion, it is important to make a point about terminology and scope right up front. Early work on subjective probability was somewhat ambiguous about whether the enterprise was intended to be of a descriptive or normative nature. Laplace is often quoted in this regard:

> The theory of probability is at bottom nothing more than good sense reduced to a calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it.

More recent authors have been careful to distinguish what our minds “feel with a sort of instinct” from what we might judge on the basis of conscious reflection and deliberation. I. J. Good, for example, distinguished ‘subjective probability’ from ‘psychological probability’ (Good, 1960). Psychological probability is “the kind of probability that can be inferred to some extent from your behavior, including your verbal communications”, while subjective probability is “psychological probability modified by the attempt to achieve consistency, when a theory of probability is used combined with mature judgment” (319-320). The distinction here is obviously not sharp. Good saw subjective probability as a refined sort of psychological probability, and in fact held that the axioms and theory of probability did not apply to psychological probability generally.

However, when I speak of subjective probability in this dissertation, I roughly mean what Good meant by ‘psychological probability’, even excluding some of what he meant by ‘subjective probability’. This, I believe, is more in line with common usage of at least some literatures, e.g., in psychology (Luce and Suppes, 1965). In the philosophical literature, Jeffrey (1965) writes, “the numerical probabilities [. . .] are taken to be subjective in the sense that they reflect the agent’s actual beliefs [. . .] irrespective of factual or moral justification” (1). Importantly, the probabilities we will use to characterize mental states will be coherent in a certain sense (see §2.7), justifying our use of the word ‘probability’.

The topic is how our minds—indeed “with a sort of instinct”, in Laplace’s words—are able automatically to encode and use probabilistic information about the world. Thus, in discussing subjective probability, we will not directly be addressing the activity of assigning numerical probabilities to various events, at least insofar as this is an explicit activity. Most of the behavior with which we will be concerned involves individual actions or decisions. For instance, just consider what our minds achieve in the course of simple activities like crossing a road. Of first importance is that we implicitly weigh the probability that a car is coming given perceptual information and general background knowledge about the context. While the implicit computations underlying this judgment may certainly involve probabilistic calculation—the Sampling Hypothesis is offered as a way of substantiating this claim—we typically do not consult the probability calculus explicitly.

To be sure, this ability can be extended and perhaps enhanced by conscious, deliberate reasoning *about* probabilities, including explicit use of the probability calculus (along with any number of other intellectual tools at our disposal). Important as this may be, our unconscious, automatic ability to assess probabilities is so much a part of our cognitive makeup that we scarcely notice it. It has often been proposed that this gives evidence that the mind somehow encodes and uses probabilistic information, which we might think of as a kind of *realism* about subjective probability. The Sampling Hypothesis is part of an attempt to support that realist claim.

2.3 Challenges to Realism

The very idea that probabilities are represented in the mind, and that we use these probabilities in making decisions and taking action, has met with considerable skepticism, both in the philosophical and psychological literatures.

To illustrate each of these worries concretely, consider the following very simple scenario, which we will also use as a running example throughout the first half of the chapter. The example is a simplified version of one from Rumelhart et al. (1986). Suppose I am describing a room in someone's house. Before I say anything about the room, how likely would you say it is to contain a chair? And how likely would you say it is to contain a stove? Whether or not you would be willing to assign numerical values to these possibilities, presumably you would judge the first to be clearly more likely than the second. Suppose I now tell you that the room contains a window and a lamp. This might make the possibility of there being a chair even more likely, while perhaps making the possibility of a stove yet less likely. If instead I had mentioned that there is a sink in the room, that might reverse things, making the stove more likely than the chair. We could go on for some time asking such questions.

These seem like relatively mundane observations. But when we inquire into the nature of the representations and processes buttressing such judgments, one obvious proposal is that we somehow encode probabilities for all of these scenarios, and our judgments and behavior are somehow based on these probabilistic representations.

The Sampling Hypothesis is a special case of this view, so it is worthwhile reviewing some of the most prominent objections and sources of skepticism toward it.

2.3.1 Complexity

Perhaps the best known objection to the idea of probabilistic representation and calculation in the mind is that it is too complex. Given the shear number of propositions, events, and situations we can entertain, many have found it implausible that the mind could have encoded numerical values corresponding to probabilities for all of them. The situation looks even worse when we consider the idea that we may need to consider arbitrary *conditional* probabilities, for either learning or inference. Many recent authors in cognitive science have expressed this worry about probabilistic models and computation (e.g., Gigerenzer and Goldstein 1996; Kwisthout et al. 2008). In the philosophical literature, Gilbert Harman famously gave voice to this skepticism in his 1986 book *Change in View*:[1]

> 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. Unhappily, this leads to a combinatorial explosion. (Harman, 1986, 25-26)

That is, even if we have only a relatively small number of "basic" propositions that are assigned probability values, if we also need to consider probabilities for these propositions conditioned on various information, that may dramatically increase the number of probabilities we need to track.

In the room example, suppose we only consider 5 possible pieces of furniture: chair, lamp, window, sink, and stove. In this case, there are 32 possible combinations of these objects, i.e., 32 possible rooms, and 80 different conditional probabilities we could ask about, e.g., whether there is likely to be a chair given that there is a window. Even in this very simple, small-scale example, it may seem farfetched to suppose that we have encoded probabilities corresponding to all of these possible

[1] See also Holton (2014) for a recent source.

queries. In Rumelhart et al.'s original scenario—only slightly more realistic, with 40 possible components of a room—there are over a trillion possible rooms, and many more conceivable conditional probabilities one ought to be able to query. As Harman (1986) concluded, subjective probabilities "are and have to be implicit rather than explicit" (36). This leaves us with the question of what that would mean.

2.3.2 Imprecision

A second criticism of probabilistic approaches to the mind concerns the idea that people's judgments under uncertainty can be quantified in such a precise way as to be captured by a unique real number value. This criticism is also common, and is nicely dramatized by Suppes (1974):

> Almost everyone who has thought about the problems of measuring beliefs in the tradition of subjective probability or Bayesian statistical procedures concedes some uneasiness with the problem of always asking for the next decimal of accuracy in the prior estimation of a probability. (160)

Worries of this nature have motivated a number of alternative formalisms, including so called *imprecise probabilities* (Good 1960, Suppes 1974, *inter alia*), designed to allow uncertainties to range within an interval, for example. Others have taken such worries to cast doubt on the usefulness of numerical representations altogether.

In the room example, it may seem outlandish to ask whether someone's subjective probability for there being a chair in a room with a window and a lamp is 0.81 or 0.81001. On the one hand, it is difficult to imagine how we could elicit such precise information in a meaningful way. On the other hand, when introspecting, people simply do not feel as though they have such definite opinions about ordinary propositions of this sort. Representation theorems can be used to assign precise values on the basis of simpler qualitative judgments, e.g., pairwise comparisons. But many have worried about what these theorems establish, given that the resulting values are still typically underdetermined (Zynda, 2000; Eriksson and Hájek, 2007). Moreover, many have criticized some of the axioms necessary for such representation theorems (again, see Good 1960 or Suppes 1974). Perhaps most importantly, a wealth of empirical research has shown that, as a point of fact, subjects routinely violate these axioms, which leads us to the third criticism.

2.3.3 Incoherence

The long line of experimental work beginning with early seminal papers by Tversky and Kahneman, e.g., (1974), has been taken by many to show definitively that people do not reason in accord with probability. Of course, the cases of explicit reasoning with and about probabilistic quantities, e.g., as in the base-rate fallacy, are consistent with the view that more basic, unconscious processes do work in a probabilistically coherent way (recall the discussion above §2.2). Unfortunately some of their results call into question the probabilistic coherence of even the most basic judgments.

In their famous conjunction fallacy (Tversky and Kahneman, 1983), subjects declare propositions of the form A *and* B to be strictly more probable than A alone. Returning once more to the room example, consider the following question. Which is more likely: that the room has a stove, or that the room has a stove and a sink? If results across many different domains are any indication, a significant number of people would affirm the latter. This is particularly bad news for anyone who wants to argue for realism about subjective probability on the basis of representation theorems. There is obviously no probability function that will agree with such an ordering, since $P(A \,\&\, B) \leq P(A)$ for any A and B. The conjunction fallacy and related empirical findings have generated a sizable literature, and I do not intend to give a new detailed diagnosis in any of what follows. I do, however, think it is important to say why these results are at least consistent with a view that takes coherent probabilistic representation very seriously.

I presume that it will be clear in what follows how these three criticisms are at least partially answered by the Sampling Hypothesis. But I will also respond to each of them explicitly.

2.4 Representing Probabilities with Propensities

The Sampling Hypothesis is based on the notion of a *probabilistic generative process.* It is common in cognitive science to assume that we are able to construct internal representations of a given domain, sometimes called *intuitive theories* or *intuitive models* (see, e.g., Gopnik et al. 2004, or Tenenbaum et al. 2011), whose primary function is to *generate instances* of some concept, event, or other in-principle-observable data associated with that domain, with different probabilities.

We can think of these generative processes as defining a random variable (or set of random variables) taking on values in some given set $\mathcal{V}$. In our toy example of the room schema, $\mathcal{V}_{\text{room}}$ might consist of the set of 32 possible room configurations, i.e., combinations of objects from among chair, lamp, window, stove, and sink. A generative model $\mathcal{M}$ for this domain would define a process that generates room instances. For instance, on one "run" of the model $\mathcal{M}$, it might generate a room with a chair, a lamp, and a window. On the next, $\mathcal{M}$ might generate a room with a stove, a sink, and a window. And so on. Each such outcome has a different probability of being generated. Thus, the generative model implicitly defines probabilities for each possible room type.

Given this, we can speak about probabilities of arbitrary events. For example, $P(\mathsf{chair})$, the probability of there being a chair in a randomly generated room, is given by the marginal probability obtained by summing over the probabilities of all those room instances with chairs. Conditional probabilities such as $P(\mathsf{stove} \mid \mathsf{sink})$—the probability of there being a stove provided there is a sink, defined in the standard way: $P(A \,|\, B) = P(A \,\&\, B)/P(B)$—would be given by the probability of generating a room with a stove, provided we only generate rooms with sinks.

The hypothesis, simply put, is that these propensities play the role of subjective probabilities. The mind encodes probabilistic information directly by its ability to support generative processes whose probabilistic dynamics mirror the probabilistic dynamics of the domain being represented. This hypothesis, in one form or another, has been suggested recently in neuroscience (Fiser et al. 2010), cognitive psychology (Vul, 2010), and developmental psychology (Denison et al. 2013)—representatives of this and other work will be discussed in what follows. While there are a number of novel aspects, new experiments and methods, and technical innovations involved in this research program, it builds on ideas from a number of different traditions in psychology and philosophy. Three particularly important precursors of the Sampling Hypothesis are: early work in psychophysics and statistical learning theory, "mental models" theories of cognition, and algorithmic work in probability and statistics. It is worth briefly reviewing these historical antecedents.

2.4.1 Historical Precedents

In psychophysics, early work by Thurstone (1927) explored probabilistic models of choice behavior that look surprisingly similar to modern sampling-based ideas, though they were quite limited in scope (cf. Luce 1959 and later developments, discussed more below in §2.10 and Appendix B). Later work on the *stimulus sampling theory of learning*, by Estes and followers, was even closer to the Sampling Hypothesis, witness the following quotation:

> It appears that the residue accumulated in a person's memory after observation of a series of events has some of the same quantitative properties as the body of data collected and processed by a survey statistician as a basis for actuarial predictions. (Estes, 1972, 82)

Suppes (1974) explicitly drew the connection between the stimulus sampling framework and subjective probability, but did not pursue the idea in any depth.

The Sampling Hypothesis, as we shall see, applies to essentially any domain we can conceive and reason about, not just directly observed statistics as in stimulus sampling. In that sense, it is more general than stimulus sampling. Especially in its application to high-level reasoning and prediction, the Sampling Hypothesis appeals to richer and more constructive "mental models" approaches to the mind, which originate from a different collection of sources.

The idea that mental objects could represent aspects of the world by virtue of resemblance is of course familiar in the case of spatial or imagistic experience, from Locke, Hume, and many others since. The suggestion behind the Sampling Hypothesis is that our basic means of representing uncertainty is also founded on a kind of structural resemblance. This idea of non-imagistic representation by resemblance was explored already by philosopher and psychologist Kenneth Craik, the forefather of "mental models" theory, who explained:

> By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the process it imitates. [. . .] Thus, the model need not resemble the real object pictorially: Kelvin's ride predictor, which consists of a number of pulleys and levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects – it combines oscillations of various frequencies so as to produce an oscillation which closely resembles in amplitude at each moment the variation in tide level at any place. (Craik, 1943, 51)

Like the tide model, a generative model has a similar relation-structure to the domain it is supposed to be a model of. The frequency of occurrences of some event represented in the mental model is supposed to mirror the assumed frequency of occurrences of the event in the world.

Thus, also like the tide model, a probabilistic generative model can be used to make *predictions* about the domain in question, on the basis of this purported resemblance. Indeed, probabilistic generative models in cognitive science are founded on the constructivist idea of *analysis-by-synthesis* (Neisser, 1967): what we internalize as cognitive agents is the ability to *construct* or *generate* instances of what we are trying to perceive, understand, or predict. The Sampling Hypothesis puts a probabilistic twist on this theme: generation is inherently stochastic.

What would a physical process implicitly representing probabilities in this way look like? An early example of a sampling machine was given by Sir Frances Galton, who used the "quincunx machine", later called the Galton box (Fig. 2.1) to illustrate various probabilistic concepts.

The probability that a given pebble will drop in a given slot is roughly proportional to the probability associated with that interval in a normal distribution. Thus, we see that after many pebbles have been dropped, with high probability the distribution

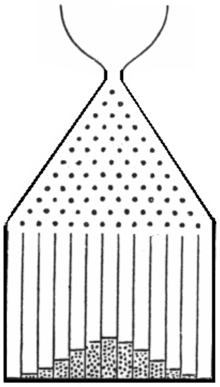

Figure 2.1: The Galton box (from Galton 1889)

resembles the normal distribution. In this picture, we can think of the different slots as the possible hypotheses—the set $\mathcal{V}$ of values—and each pebble as a sample from the underlying distribution over $\mathcal{V}$.

Of course, this is meant only as an illustration. More serious proposals have been inspired by work in computational probability and statistics, where efficient methods for sampling from complex distributions has been extensively explored (see §2.6-2.7 below). Intriguingly, some of this work has proceeded in tandem with psychological investigations in the connectionist tradition. While probabilistic and connectionist approaches to the mind are sometimes viewed as opposed to one another, some of the most suggestive examples of concrete sampling processes come from work in connectionism. In the next section, we consider one such example.

2.5 Sampling Machines

The room example can be modeled using a *Boltzmann Machine* (Rumelhart et al., 1986), which is a neural network $\mathcal{N} = \langle N, \mathbf{W} \rangle$ given by:

- a set N of binary nodes, in this case $N_{\text{room}} = \{\mathsf{chair}, \mathsf{lamp}, \mathsf{window}, \mathsf{stove}, \mathsf{sink}\}$;
- a weight function $\mathbf{W} : N \times N \to \mathbb{R}$, s.t. $W_{i,i} = 0$ and $W_{i,j} = W_{j,i}$ for all $i, j \in N$.

The particular example we have been discussing is depicted in Figure 2.2.

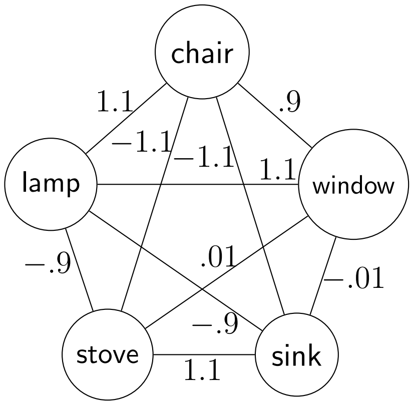

Figure 2.2: Boltzmann Machine $\mathcal{N}_{\text{room}}$ for the simple room schema

Intuitively, the weights between nodes represent the (positive or negative) correlation between what the nodes represent. Thus, stoves and chairs are anti-correlated, while stoves and sinks are correlated. The basic *activation function* for the Boltzmann machine is a stochastic update rule that determines whether a node i will be on or off, as a function of whether other nodes are currently on or off and the weights between those nodes and i. In particular, a node i is randomly chosen, and is turned on (or remains on) with probability given by the familiar logistic function:

$$\frac{1}{1+e^{-net_i}}$$

where $net_i = \sum_j W_{i,j}\mathbb{I}_j$, and $\mathbb{I}_j$ is an indicator function, equal to 1 if x_j is currently activated, and 0 otherwise.

As explained in Appendix A.1, this simple activation rule can be seen as carrying out the so called *Gibbs Sampling* algorithm on an underlying distribution given by an associated energy function. This *Boltzmann distribution* gives us a well-defined probability P_{room} on sample space $\mathcal{V}_{\text{room}}$, which we can use to model the judgments discussed above. With the above weights we can calculate that

$$P_{\text{room}}(\mathsf{chair} \mid \mathsf{window},\ \mathsf{lamp}) = 0.81 \quad \text{while} \quad P_{\text{room}}(\mathsf{chair} \mid \mathsf{stove},\ \mathsf{sink}) = 0.24.$$

Furthermore, while the prior probability of there being a stove is low (0.30), provided there is a sink the probability of there being a stove as well is above chance (0.56).

The details of this particular example, and of the numbers given, are not important. There are two main ideas I want to convey with it. The first is that it shows one way that a relatively complex distribution, including a great deal of information about *conditional* probabilities, can be encoded with an extremely compact representation. The network depicted in Figure 2.2 is quite small compared to the number of (conditional) probabilities it implicitly defines.

Second, and most importantly, $\mathcal{N}_{\text{room}}$ represents the distribution P_{room} precisely in the sense that the probability of this machine outputting a given value $v \in \mathcal{V}_{\text{room}}$ is $P_{\text{room}}(v)$. That is, if we apply the activation function as described above for a sufficient number of steps, the resulting vector of values will amount to a *sample* from distribution P_{room}. Likewise, if we do this with some of the nodes "clamped" to remain on, then the network will sample from the corresponding conditional distribution. E.g., clamping window and lamp, we can use the machine to sample from $P_{\text{room}}(\mathsf{chair} \mid \mathsf{window},\ \mathsf{lamp})$. Indeed, given how Boltzmann Machines and related tools have been used in cognitive modeling, we would expect sampling in this way to be the primary means of extracting probabilistic information from the underlying distribution. The exact calculations above are quite involved and had to be done by computer. They do not describe information that is itself readily available to some other mechanism which might use this machine. Instead, these probabilities describe information that is implicit in the network, and which can be approximated by running the network in the manner described.

To a certain extent, these observations already show how we meet the complexity challenge raised by Harman and others (§2.3.1). It is by making probabilities implicitly, rather than explicitly, represented, just as Harman suggested.[2] The network $\mathcal{N}_{\text{room}}$ defines all 80 conditional probabilities associated with this scenario in an extremely compact format. In fact, there are formal results to this effect, showing that

[2] Though this is a very different proposal from what Harman suggested (Harman, 1986, Ch. 1). See Millgram (1991) for a response to Harman also drawing on (approximate inference for) graphical models. Millgram goes further, questioning whether any "hardness argument" could ever cast doubt on the idea that we *ought* to use coherent probabilities in reasoning.

the *efficiently samplable* class of distributions, under a suitable formalization, form a proper superset of the class of *efficiently computable* distributions, provided P $\neq$ NP (and weaker assumptions suffice, see Ben-David et al. 1992).

2.6 Some Clarifications

Before proceedings further, I would like to comment on the scope of the Sampling Hypothesis as explained so far: concerning the types of distributions represented, the mode and medium of representation, and how these implicitly defined probabilities might actually be used.

The Boltzmann Machine described above can only represent certain kinds of distributions. In the parlance of graphical models, they are *Markov random fields.* These are essentially undirected models in the sense that all probabilistic dependencies are assumed to be symmetric. There are many other graphical models that have been used in cognitive science and artificial intelligence (Pearl, 1988; Koller and Friedman, 2009), of which the best known may be Bayesian networks. All of these models have in common the idea of reducing the amount of information from the full probability table that needs to be encoded, taking advantage of (or just assuming) independencies between variables. They do this in different ways. Bayes nets, for example, allow asymmetric dependencies, but unlike Markov random fields may not contain cycles. The distinctions among these models are not important for our purposes, and we could just as well have used any graphical model with an appropriate sampling algorithm to illustrate the main points.

In fact, it has become increasingly common in these fields to use formalisms more expressive than traditional graphical models. For instance, Markov Logic generalizes the Boltzmann Machine and Markov random fields by allowing relational statements and unbounded state spaces (Domingos and Lowd, 2009). Probabilistic grammars, which also define probabilities over infinite spaces, have proven useful in modeling aspects of language understanding, causal reasoning, and concept learning (see Tenenbaum et al. 2011 for an overview and many references). Some have even suggested cognitive science might avail itself of *universal* formalisms and languages, employing unrestricted recursion, in which any computable distribution can be represented (Goodman et al., 2008; Freer et al., 2012). In this chapter, we can remain neutral about exactly what kinds of computational processes underly our internal generative models, and how much logical structure we need to capture those capacities. Of course, to the extent that a formalism can be implemented in a neurally inspired physical mechanism, that speaks in its favor. In this sense, the Boltzmann Machine is a suggestive example, but we will also see other examples (§2.8).

Independent of the abstract computational form of these generative processes, there is also an obvious question of how the Sampling Hypothesis relates to debates about where and how computations are performed in the brain. There is suggestive evidence that the mind uses episodic memory traces to simulate possible future events for the purpose of prediction (Schacter et al. 2008, who suggest this may even be the primary function of episodic memory). This is certainly compatible with the Sampling Hypothesis. It should also be clear that the hypothesis is compatible with views according to which reasoning and prediction are grounded in perceptual/motor processes (Barsalou, 1999). In general, there may or may not be a sense in which the propensity of a simulation mechanism resulting in a particular outcome has anything to do with the agent's subjective probability of that outcome. Conversely, it may be that the generative mechanisms underlying some cognitive capacities are made up of more abstract "symbolic" representations. In other words, we can also be neutral about what the elements of $\mathcal{V}$—the output values of the generative model, i.e., what is being simulated, as well as the other elements of the generative process—in fact are, and how they are realized. The same holds for other related debates, e.g., about distributed vs. localist computation, as the Boltzmann Machine example makes clear. Connections between computational models and neural implementation will be discussed briefly in §2.8, but it is important to stress that the hypothesis is not anchored to any specific proposal.

Next, we have not yet said anything precise about how these representations of uncertainty are used by a subject. Here is an example of a decision rule illustrating how samples could be used for a simple elicitation task, which we will use (sometimes without mention) throughout this chapter:[3]

> Decision Rule A: Given a task to report which of n exclusive and exhaustive hypotheses $\mathcal{V} = \{H_1, \dots, H_n\}$ is true, and a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values, take R samples from $\mathcal{M}$ and let best be the set of hypotheses that receive the largest number of samples. Return $H \in \textsc{best}$ with probability $\frac{1}{|\textsc{best}|}$.

There will obviously be such a rule for any number R of samples. For instance, using the Boltzmann Machine in Fig. 2.2 to answer the question, "Will a room with a stove also have a sink?", Decision Rule A with $R = 3$ would have us clamp the stove node and run the network for a while before seeing whether the sink node is actived, repeating this three times.[4] With very high probability, at least two of those three runs would have sink actived, so Decision Rule A would likely return "yes".

One can imagine any number of variations and elaborations of this simple decision rule (see §2.11). What is important is that the propensities associated with these mechanisms can reasonably be interpreted as representations of uncertainty, and that the outputs can therefore be seen as genuine samples from a meaningful distribution. No one doubts the mind's ability to come up with instances of a concept, to imagine possible future events, and so on. It is moreover obvious that on any given occasion we could associate a probability distribution with the possible outcomes of these processes. The substantive claim is that these sampling propensities can be understood as encoding the subject's own uncertainty. This will presumably be based on some combination of how these processes relate to the agent's experience and observations, and especially how they are used in producing verbal and other behavior, using a decision rule similar to that given above.

How wide-ranging is the Sampling Hypothesis supposed to be? Are we to suppose that every judgment under uncertainty is the result of a sampling process using some internal generative model, invoking something like Decision Rule A? Such

[3] This only works for relatively small discrete hypothesis spaces. For continuous spaces a natural alternative would be to construct a density estimate and return the mean, for instance.

[4] For the Boltzmann Machine, which defines a *biased* sampler (see §2.7, §2.9, Appendix A.1), one must specify how many iterations to run before the current state is returned as a *bona fide* sample.

an assumption would obviously be too strong. As we know, people do occasionally make such judgments using explicit reasoning and calculation, and there may be other unconscious strategies that the mind employs. The full extent of sampling in human cognition is of course an empirical question, so it will be more illuminating to let the empirical evidence speak for itself.

First, however, I would like to make one final clarification concerning the relation of the Sampling Hypothesis to the Bayesian program in psychology.

2.7 Sampling and Bayes

The Sampling Hypothesis has been stated most explicitly in the literature on Bayesian psychology. There is good reason for this. Probabilistic models of psychological phenomena are often quite complex, and computing with conditional distributions can often be intractable. For instance, in Bayes Rule, computing the normalizing constant $Z = \sum_{H' \in \mathcal{V}} P(E|H')P(H')$ can be very costly if there is a very large number of possible hypotheses in $\mathcal{V}$:

$$P(H|E) = \frac{1}{Z}\, P(E|H)P(H)\ .$$

In place of exact calculation, it is sometimes possible to approximate probabilistic computations using samples from the distribution, which may be relatively easy to produce. Suppose, for instance, one needs to calculate an expectation Φ of a function $\phi(X)$ under distribution $P(X)$,

$$\Phi = \int P(X)\phi(X)\ dX\ .$$

For instance, in an expected utility calculation for an action A, we would have $\phi(X) = u(A, X)$, though there are many other examples. Provided one can effectively generate samples $\{X^{(r)}\}_{r=1}^{R}$ from $P(X)$, the expectation Φ can be approximated by

$$\hat{\Phi} = \frac{1}{R}\sum_{r} \phi(X^{(r)})\ .$$

The Law of Large Numbers guarantees $\hat{\Phi}$ will converge to Φ as R goes to infinity. Moreover, the variance of $\hat{\Phi}$ decreases as R increases, which means the more samples taken, the closer the estimate is expected to be to the target distribution. Recall the Galton box (Fig. 2.1), where this can be visualized.

Computational uses of probability have led to efficient algorithms for approximate sampling, collectively called *Monte Carlo methods*, of which Gibbs sampling is a prime example (MacKay, 2003, Ch. 29). Since sampling exactly from a distribution is typically as hard as exact calculation, these methods produce biased samples, in which the sequence of random variables is correlated. It can be shown that this bias washes out after enough iterations of the procedure.[5]

While a connectionist modeler might use a Boltzmann Machine and note incidentally that it embodies an underlying probability distribution over states, the Bayesian modeler typically begins with the probability distribution before exploring more concrete algorithms. Several recent authors in this literature have proposed that Monte Carlo methods and related stochastic sampling algorithms provide *algorithmic level* hypotheses corresponding to the *computational level* models at which ideal Bayesian analyses are targeted, in Marr's sense of these terms.[6] They allow people to approximate ideal Bayesian inferences and decisions using a large enough number of samples as proxy for the explicit calculations. Indeed, some of the most compelling experimental demonstrations of the viability of the Sampling Hypothesis stem from these proposals (see §2.8-§2.10), which additionally purport to show the distributions from which people are effectively sampling are rational, e.g., by virtue of being appropriately conditioned on observed data.

[5] In fact, these biases have themselves been used to explain certain aspects of behavior, as reviewed below in §2.8 and §2.9.

[6] See Marr (1982) and Chapter 3 of this dissertation. Compare this with the following quotation from Churchland and Sejnowski (1994):

> In the Boltzmann Machine, matters of computation, of algorithm, and of implementation are not readily separable. It is the very physical configuration of the input that directly encodes the computational problem, and the algorithm is nothing other than the very process whereby the physical system settles into the solution. (92)

For the Bayesian psychologist, the mere specification of the machine leaves out a crucial level of computational *explanation*: understanding what the machine is computing.

What is the relation between the Bayesian program and the Sampling Hypothesis? Let us take the Bayesian program (the descriptive, not prescriptive, version) to consist in two claims:

(1) Mental states can be described in terms of (coherent) probability distributions;

(2) Learning and updating amount to conditioning an appropriate distribution.

The Sampling Hypothesis is committed to some limited version of (1), but it is reasonably neutral with respect to (2).

Concerning (1), in some sense coherence is built right into the Sampling Hypothesis. A generative model $\mathcal{M}$ implicitly defines a distribution on a sample space $\mathcal{V}$, and we can certainly extend this to a coherent distribution over a full event space. We were already doing this in §2.5 when talking about conditional probabilities of complex events in the room model. Such a model, on the Sampling Hypothesis, characterizes knowledge implicit in a subject's mind, though by itself it tells us nothing about how a subject will use the model. For that we need something like the DECISION RULE A. From the fact that sampling is inherently noisy, we might expect a subject using such a rule to give inconsistent responses to the same question on different occasions, even when knowledge of the domain has not changed. This is indeed what we find (§2.10). Thus, characterizing a concrete psychological mechanism *from outside* using a coherent probability distribution certainly does not rule out the possibility that we would want to characterize the agent's *behavior* as incoherent or inconsistent.[7] Additional sources of incoherence stem from the possibility that a subject will construct two different models to answer the same question, e.g., on account of framing effects, or more generally will construct the "wrong" internal model, or use the right model in the "wrong" way. It is tempting to suggest that something along these lines is behind the conjunction fallacy.

[7] Cf. Luce and Suppes (1965). It might seem a stretch to call these probabilities the agent's *beliefs*. We evidently cannot determine "by introspection" what these sampling propensities are (though we may be able to learn something about them in this way). Rather, they are (partially) manifested in an agent's behavior, and as theorists we hypothesize these sampling propensities only on the basis of extensive empirical investigation and systemization (§2.12). For these reasons, I have deliberately avoided calling these probabilities 'degrees of belief'.

As for (2), suppose we have established that an agent's prior probability in some situation is given by a generative model corresponding to distribution $P(\cdot)$. Then upon receiving some information E, (2) says that the agent's new model of the situation should correspond to distribution $P(\cdot\,|E)$. In the Boltzmann Machine this required only a simple adjustment to the network: clamping the nodes corresponding to E. And to repeat, much psychological work suggests that in many cases subjects do condition according to Bayes Rule. At the same time, few would venture that people are ideal Bayesian learners, with their current state of knowledge the result of a lifetime of perfect updates to an initial prior probability distribution (though see Perfors 2012 for a seemingly optimistic view). The use of some Monte Carlo methods, such as the particle filter (see §2.9 below, and Chapter 3), assumes that subjects will settle on *locally approximately optimal* hypotheses, and use those as assumptions in future inferences rather than always reconsidering whether those earlier hypotheses are still supported by the data, to take just one example. Other authors have also stressed the local nature of prediction and judgment, and the prevalent flouting of Carnap's Principle of Total Evidence, according to which judgments should be made taking into account *all* of the evidence available (see, e.g., Gaifman 2004). The full extent to which people are Bayesian in the sense of (2) is to a large extent an empirical question, one to which we will return in Chapter 3. The Sampling Hypothesis by itself is consistent with a range of possible answers to this question.

2.8 The Neural Basis of Subjective Probability

The neural basis of subjective probability is a matter of some controversy in contemporary neuroscience (Knill and Pouget, 2004; Vilares and Kording, 2011). It is generally accepted that neural firings are inherently noisy, and one of the central questions is how this noise might be used to the brain's advantage. Noise is suggestive of something like sampling, but there are several alternative proposals. Perhaps the most prominent conjecture is the *coding hypothesis*, according to which spike rates of neurons directly represent probabilistic numerical quantities, such as likelihood ratios or parameters of density functions. Firing rates have been shown to represent other continuous quantities such as direction of movement, and probability is assumed to be simply another type of quantity that can be represented in a similar way. Sophisticated proposals show how an appropriate kind of noise conjectured to characterize neural firings can aid in decoding (see Knill and Pouget 2004 for a review).

The Coding Hypothesis has recently been partially corroborated by Yang and Shadlen (2007), who demonstrated that the firing rates of certain neurons in the monkey cortex can be adequately modeled by (a linear function of) the logarithm of the posterior odds ratio between two hypotheses H_1 and H_2 given some evidence E:[8]

$$\log \frac{P(H_1|E)}{P(H_2|E)}$$

The monkeys are presented with a series of shapes, each of which is associated with a randomly pre-assigned weight. At the end of a trial, the monkey saccades left or right and is rewarded with probability proportional to the summed weights for the hypothesis chosen. Strikingly, following the training phase, the monkeys' neural firing rates are successively updated after each new shape is presented, accurately reflecting the posterior probabilities, which in turn predict the monkeys' responses.

It is worth pointing out that, formally speaking, the neurons in this study could easily fit into a larger Boltzmann-like network. Recall that in the Boltzmann Machine the probability of an individual neuron x_i firing, given the current state of the other neurons x_j, is the result of applying the logistic function to the net input of neuron i:

$$\frac{1}{1+e^{-net_i}} \, .$$

It so happens that under the Boltzmann distribution, net_i is equal to the log odds ratio (see Appendix A.1). The results from Yang and Shadlen (2007) suggest that instead of applying the non-linear logistic function to net_i, the activity in these neurons is governed by a simpler, linear function of its input. This makes sense if these neurons are simply serving as summaries of computations performed elsewhere—perhaps by

[8] The results in the paper are stated in terms of the log likelihood ratio, also known as the *weight of evidence* (Good, 1983); but since the two hypotheses have equal prior probability in these experiments, these values are equivalent by Bayes Theorem.

sampling—to be used directly for action selection. After all, as the authors mention, it is an open question where and how these computations occur. They were only able to measure the results of the computations. To that extent, it could very well be that there are roles for both coding and sampling in how the brain manages uncertainty.

Indeed, several other authors have pointed out ways coding and sampling might work in conjunction. As Lochmann and Deneve (2011) note, in cases with only two possible states, as in the Yang and Shadlen experiment, spike rates can themselves be interpreted as rapid samples from a distribution. Working in the coding framework, Moreno-Bote et al. (2011) present a sophisticated account of how the brain could sample directly from distributions represented by *population codes*, one of the most mathematically well-developed versions of the coding hypothesis. Sampling is at least consistent with what we know about neural representation. But is there any reason to suppose it is actually used by the brain?

Most of the work on neural coding of probability focuses on low-level perceptual tasks, where the relevant distributions are relatively low-dimensional and can typically be described in terms of parametric models, e.g., using normal distributions. In these cases it is possible to summarize the relevant statistics by only a few values, e.g., mean and variance. However, as many have noted (Fiser et al., 2010; Vul, 2010), coding models do not easily scale up to more combinatorially complex tasks such as categorization, language understanding, causal inference, or higher-level visual processes. For such tasks, sampling is a favored engineering solution (MacKay, 2003). It is tempting to postulate that the brain may have happened upon similar solutions. In fact, there is behavioral and modeling evidence that this may well be the case.

One of the most intriguing empirical arguments for sampling comes from *multistability phenomena* in perception, e.g., as with ambiguous visual images, where subjects report randomly alternating between interpretations. A special case of multistability is binocular rivalry, where one eye is presented with one image and the other eye with another (Figure 2.3). In these cases subjects report percepts flipping back and forth, occasionally resting briefly in intermediate "hybrid" states.

One feature of the Boltzmann Machine with Gibbs sampling (and many related models) is that, even after it has settled into a steady state, with positive probability

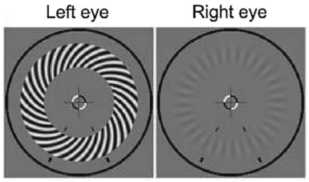

Figure 2.3: Binocular rivalry experimental stimuli from Gershman et al. (2012)

it will transition to another state. If there are several energy minima in the state space (states with high probability), such transitions will occur with high probability. A number of researchers have noticed that this suggests a possible analysis of multistability phenomena. Indeed, it is common to view perception as the brain's solution to an inference problem, namely guessing the latent state of the world from noisy perceptual data (an idea going back to Helmholtz). To the extent that different latent states have different probabilities given the perceptual input, the Sampling Hypothesis predicts that the perceived image will be some function of a sample (or samples) from this distribution on interpretations. One of the attractive aspects of the multistability phenomena is that we can observe the behavior of the visual system over an extended period with the same input.

In one recent paper, Gershman et al. (2012) define a probabilistic model for inferring luminance profiles from two separate retinal images—in fact based on a variation of the Boltzmann Machine—and show that the Gibbs sampling algorithm predicts a number of central empirical results from the literature on binocular rivalry and multistability.[9] For instance, they are able to explain the gamma-like distribution describing the time between switches, which had required *ad hoc* postulations in previous accounts of multistability. On their model, this is a natural consequence of the sampling dynamics and the network topology, as are several other phenomena associated with multistability.

Remarkably, a concrete implementation of the Sampling Hypothesis accounts for

[9] They build on a large literature on this topic, including earlier probabilistic and sampling-based analyses of multistability, especially Sundareswara and Schrater (2007). See Gershman et al. (2012) for other references. See also Moreno-Bote et al. (2011), cited above, for a different sampling analysis based on population coding.

many of the detailed experimental findings on this phenomenon. It suggests that in multistable perception we are actually witnessing sequences of samples as conscious visual images.

There are some features of the Boltzmann Machine, and Gibbs sampling in particular, that make it less plausible as a neural model. For instance, the underlying dynamics of Gibbs sampling is *reversible*, in the sense that the probability of starting in state s_1 and transitioning to s_2 is the same as that of starting from s_2 and transitioning to s_1. Networks of spiking neurons in the brain do not seem to have this property, e.g., because of the refractory period following a spike. Buesing et al. (2011) propose an alternative, more biologically realistic sampling mechanism with non-reversible dynamics, which they show converges quickly and predicts empirical spike timing data. They also show that their alternative algorithm can account for many of the same multistability phenomena as in Gershman et al. (2012). Further work extends the framework to the continuous setting, and to more general graphical models. The details of this work are beyond the scope of this chapter, but it represents one promising line of research bridging concrete generative models like the Boltzmann Machine with details of what we know about neural processing.

Sampling has also been implicated in higher-level cognitive phenomena involving logically complex spaces, such as categorization, language understanding, causal cognition, and many others,[10] areas where alternative computational models that work in low-level vision seem ill-suited.

To give a concrete example, consider the problem of parsing a sentence. It is common to assume that sentences have some underlying structure and that the task is to infer the appropriate structure from the surface form of the sentence. In general, there are many possible structures correspond to any given surface form, making parsing a hard problem. Levy et al. (2009) show that a model based on the *particle filter*—a Monte Carlo algorithm designed for sequential addition of data—is able to predict

[10] As we are focusing on inference here, rather than learning, we put off a discussion of category learning to the next chapter. Sanborn et al. (2010) analyze categorization using particle filters as an approximation to the proposed ideal Bayesian calculation, and demonstrate a close fit to the data. Like with parsing, discussed in the next paragraph, particle filters with relatively few particles crucially predict *order effects* characteristic of human categorization behavior. See also Denison et al. (2013) for a recent application of a very simple sampling algorithm to causal learning in children.

some of the central findings from the literature. The basic idea is that a number of "particles" are maintained with different weights, corresponding to hypotheses about the correct interpretation of the sentence. As each word is perceived, a new set of (weighted) particles is drawn, probabilistically depending on the new word and the particles (and their weights) from the previous step. Among other predictions, the particle filter algorithm accounts for "garden path" sentences, like that in ($*$) below, which subjects routinely take a long time to process.

($*$) The woman brought the sandwich from the kitchen tripped.

If the number of particles is too small, interpretations that seem to have low probability initially—here, the interpretation on which 'the woman' is the object of 'brought'—may simply drop out, and the algorithm "crashes" upon processing 'tripped', requiring the analysis to start over. With the number of particles intuitively standing in for amount of working memory, Levy et al. (2009) were able to show a good fit to experimental data.

2.9 Heuristics and Biases

So far we have seen that sampling is implicated in relatively automatic psychological processes like vision and parsing. One might wonder to what extent high-level judgments and predictions are the product of a sampling process. That people do not always behave like ideal probability theorists in this kind of task is by now commonplace (§2.3). The large body of work by Tversky and Kahneman (1974), and much work following, has identified a number of general *heuristics and biases* that people seem to use, and their effects are by now well established. This program has sometimes been characterized as an alternative to probabilistic models, and especially "rational" probabilistic models (e.g., Gigerenzer and Goldstein 1996, *inter alia*). However, an interesting twist explored in very recent work is the idea that some of these heuristics and biases can be seen as natural consequences of sampling from appropriate probability distributions.[11]

[11] Much of the interest from the work described below stems from the fact that this behavior can be "rationalized" in terms of optimal resource/accuracy tradeoff. This will be discussed more in the

One of the classic examples discussed in Tversky and Kahneman (1974) is the *anchoring and adjustment heuristic.* In one of their experiments, they asked subjects to estimate some unknown quantity n, e.g., the percentage of African countries in the United Nations, after first rolling a roulette wheel with numbers 1-100 and having subjects say whether n is greater or less than the number on the wheel. The finding, which has been reproduced many times and in many different contexts, is that responses are highly skewed toward the number on the roulette wheel, suggesting that people take that number as an initial hypothesis ("anchor") and "adjust" it to a more reasonable value. For instance, on the UN question, the median response from subjects who saw a 10 on the wheel was 25; it was 45 for the group that saw 65.

On the face of it, the anchoring and adjustment idea has a similar structure to Monte Carlo algorithms. The latter always begin in some initial state, i.e., with an initial hypothesis, and then "adjust" the hypothesis as the search space is explored along the constructed Markov chain. Recall the Boltzmann Machine. The network begins with some initial activation vector, with each node on or off, before stochastically updating nodes according to the logistic activation function. Clearly, if the network is only run for a brief amount of time, it will likely remain close in state space to the initial vector. This is a general feature of many Monte Carlo algorithms.

In recent work, Lieder et al. (2012) have demonstrated that much of the data on anchoring and adjustment can indeed be accurately modeled by assuming people are sampling from an appropriate posterior distribution and taking some salient answer as initial value.[12] Moreover, they make the critical point that viewing this behavior through the lens of sampling suggests a more general perspective. Whereas most of the work on anchoring and adjustment is based on point/number estimation problems, which have an obvious distance metric on possible answers, this can be seen as just a special case of more general inference problems, including over rich hypothesis spaces, with a more general notion of "distance" between possible responses. At any rate, the very fact that subjects exhibit this bias can be interpreted as further confirmation of the Sampling Hypothesis, and specifically that the mind uses something like these

next chapter.

[12] For details, see Lieder et al. (2012). As noted above, they furthermore take this to be a *rationalization* of the heuristic (for further discussion see §3.6).

particular Monte Carlo algorithms.

Other well known heuristics from this literature also seem consistent with sampling. Whenever these generative models must be *constructed* on the spot, e.g., from fragments of memory, we should expect any biases associated with memory retrieval to affect predictions. Another of Tversky & Kahneman's (1974) examples that illustrates this is the *availability heuristic*. In one of their original experiments, subjects were asked to estimate how likely a word in English is to end in 'ing' and (separately) how likely a word is to have 'n' as the penultimate letter. They judged the former much more probable than the latter, even though the latter are a strict subset of the former. (Notice this is another instance of the conjunction fallacy mentioned in §2.3 above.) The received explanation is that it is much easier to probe one's memory for 'ing' words. So when a subject tries to determine how probable such words are, assuming this is done by taking a random sample of instances from memory, a larger number of examples will be brought to mind when assessing 'ing' words than when assessing penultimate-'n' words.[13] Thus, if a subject were using a variation on Decision Rule A to estimate these two probabilities, we might expect such a result.

Rather than casting doubt on the Sampling Hypothesis, many of the observed heuristics in judgment and prediction behavior corroborate the idea that people deal with such tasks by sampling from internally constructed models. Understanding how such models are constructed and how samples are generated are important ongoing research questions. Work such as that of Lieder et al. (2012) provides hope that specific sampling algorithms may give important insight into these questions. It moreover strengthens the suggestion that these samples are drawn from distributions that we can sensibly view as internally representing the subject's own uncertainty.

[13] The literature following Tversky and Kahneman (1974) investigated an ambiguity in the statement of the heuristic, as to whether probability judgments are based on the *number* of instances brought to mind, or the *perceived ease* with which examples are brought to mind. Experimental results by Schwarz et al. (1991) suggest it is the latter, which foreshadows a complicating factor in the story: monitoring one's own use of an intuitive model and taking results thereof as evidence. Metacognition and metareasoning will be the main topic of Chapter 4.

2.10 "The Crowd Within"

The various sampling algorithms—Gibbs sampling, particle filters, etc.—each have their own characteristic biases and dynamics. As explained in the previous sections, these dynamics can often be shown to match aspects of empirical psychological data. What all of these algorithms have in common is that they can account for approximate *probability matching* behavior.

Suppose in some experiment subjects are presented with two alternative event types A and B on 70% and 30% of trials, respectively. Population-level probability matching occurs when roughly 70% of subjects respond with A and 30% with B. That is, instead of each subject always choosing the most frequently observed event type, the distribution of responses matches the empirical distribution associated with the stimuli. As Vul et al. (2013) and others have pointed out, this extends beyond simple frequency or probability matching to, so to speak, *posterior matching*, where the distribution of responses actually matches some normative posterior distribution for the task under investigation. This has been observed across a number of domains, including categorization, word learning, causal learning, psychophysics, and others.

The simple Decision Rule A from §2.6 with $R = 1$ sample (and in fact also with $R = 2$) predicts probability matching exactly. The probability of responding with H is given by the probability of drawing H as a sample, which just is the probability of H. Thus, sampling can account for probability matching by supposing that subjects are making decisions in these cases by drawing only one or two samples. In fact, Vul et al. (2013) have even shown that there is a correlation between the stakes of a given problem and the number of samples that would explain subjects' behavior if they are using something like Decision Rule A. People can, and apparently do, strategically adjust the number of samples they draw before giving a response, depending on how much of a difference they could expect more samples, and thus presumably better accuracy, to make.[14]

It is worth pointing out that in some ways this explanation of probability matching makes more sense than the standard account, according to which subjects are using

[14] This again introduces complications related to metareasoning, which one might think should be modeled as well. We will treat this very phenomenon of strategic sampling in the next two chapters.

a *softmax decision rule* (sometimes called the *generalized Luce-Shepard rule*). In the specific case of estimation problems—where utility of an estimate is 1 if correct, 0 otherwise; thus expected utility and probability coincide—the softmax rule says that a subject will respond H with probability

$$\frac{e^{v(H)/\beta}}{\sum_{H' \in \mathcal{V}} e^{v(H')/\beta}}$$

where $v(H) = \log P(H)$. For $\beta = 1$, this is equivalent to Decision Rule A with $R = 1$ (or 2), i.e., probability matching. But like Decision Rule A, the softmax rule can model the gradient between probability matching and maximizing by varying the value of β. As β goes to 0 the probability goes to 1 for the most probable hypothesis.

While the two can be used almost interchangeably to explain probability matching behavior, a literal application of the softmax rule requires subjects to be able to compute (e.g., posterior) probabilities perfectly. The noise is interpreted either as encoding our own uncertainty about utility (McFadden, 1973), or perhaps as random error in action execution. It seems clear that Luce (1959) intended the model merely as an adequate *description* of choice behavior (see also Luce and Suppes 1965), and it has since been used frequently across a number of psychological domains because of its flexibility in fitting data. Sampling-based decision rules like Decision Rule A may offer a principled, more mechanistic explanation of this behavior (Vul, 2010; Vul et al., 2013), at least in some cases. Often it seems as though the difficult aspect of a decision problem—where the brain may need to take a shortcut—is precisely in estimating what will happen under various scenarios. A more extended discussion of this issue can be found in Appendix A.2.

Decision Rule A attributes population-level variability to the inherently stochastic nature of individual choice. But probability matching at a population level is of course consistent with each individual subject following some deterministic procedure. This potential objection is especially pressing given that probability matching behavior has been shown to evolve in simulated populations of completely deterministic agents (Seth, 1999).[15] Why in a given case should we suppose subjects have internalized roughly one and the same generative model and are each drawing only a few samples from it, when we may be able to explain the same aggregate behavior by, say, individual variation? (Cf. Mozer et al. 2008.)

If subjects' responses are drawn from a distribution associated with an internal generative model, we would ideally like to elicit multiple responses from the same individual and use those to estimate various statistics of that subject's assumed distribution. This would give the most direct possible behavioral evidence for the Sampling Hypothesis, and is essentially what we seem to have in the case of binocular rivalry. However, for cases of ordinary reasoning and prediction, this is in practice complicated by the fact that subjects might remember their earlier responses and not resample. Intriguingly, earlier experimental investigations of choice behavior showed that subjects would nonetheless offer inconsistent responses when probed with the same question, if separated by many other questions during which time they would presumably forget their earlier answers (for a review see Luce and Suppes 1965, §5).

More recently, Vul and Pashler (2008) performed a similar study, specifically on point estimation problems like the UN example above (§2.9), in order to test whether averaging multiple guesses by the same subject would be more accurate than the average accuracy of their guesses, as has been demonstrated for groups of subjects. They found this to be the case. Furthermore, averages were more accurate for subjects tested three weeks apart than twice on the same day, suggesting that samples become more independent as time goes on. Thus, not only do subjects exhibit random variation, but responses appear to originate from some underlying distribution, which itself may encode more accurate knowledge of the world than any single sample drawn from it. In some sense, this latter point is obvious. When making an estimate, say, about the percentage of African countries in the UN, we bring to bear all sorts of knowledge about Africa, the UN, and any other relevant topic, which cannot be

[15] Moreover, in some psychological cases it appears that matching behavior is the result of individual variation in nearly-deterministic rules. See, e.g., Goodman et al. (2008) or Sanborn et al. (2010). The results there are consistent with subjects *learning* concepts by sampling, viz. hypothesis sampling, though they seem to make *inferences* on the basis of one or two learned rules.

captured by a single numerical estimate. What is interesting about Vul and Pashler's results is that repeated elicitation of estimates gives evidence that subjects' "intuitive theory" of many of these domains is surprisingly accurate, and that this intuitive theory can be used to produce samples from a sensible distribution, as the Sampling Hypothesis proposes.

2.11 Representing Value

The reader may have noticed conspicuously little discussion so far of utility, preference, desire, goals, or other value-based representations and attitudes. This may seem especially unscrupulous given that one of the main difficulties with traditional representation-theorem methods for operationalizing subjective probability is in the combination of probability and utility. As noted in §3.1, one perennial problem is that behavior and choice seem to underdetermine representations of probability and value and how they are combined. Have we skirted this issue at our peril? Surely the results discussed in the previous three sections require making some assumption about what subjects value, or about what a given mechanism is aiming toward.

This omission is in fact deliberate, for several reasons. First, it seems safe to assume that the (ordinal) utility structure in many of the problems discussed so far is essentially trivial. For instance, in an estimation problem one's prediction is either correct or incorrect. In many such cases, we can safely assume expected utility is essentially equivalent to probability (perhaps scaled by some factor), so there would be little point for the brain to represent utility explicitly. This is especially pertinent in automatic, relatively encapsulated domains like vision and parsing, where the utility structure may be fixed. In these cases, one might argue that we are afforded a glimpse into how the mind represents probability without having to worry much about explicit representation of value.

Second, there is much current debate about whether it is even possible to disentangle representations of probability and value in the brain (for a recent review and discussion, see Gershman and Daw 2012). To be sure, the examples where utility is relatively unimportant (e.g., in the study by Yang and Shadlen 2007) do not show that we can separate the two, since by assumption expected utility and probability are indistinguishable. Instead of taking a definite stance on this controversial issue, the aim has been to present sufficient evidence for the Sampling Hypothesis with as little dependence on utility assumptions as possible. This neutrality is arguably justified by the observation that the Sampling Hypothesis is, in an interesting way, compatible with quite different assumptions about how value figures into decisions.

Consider a more general version of our Decision Rule A from §2.6, distinguishing the state space from the action space, and defining an explicit utility function:

> Decision Rule B: Suppose we are given a task with possible states $\mathcal{V} = \{H_1, \ldots, H_n\}$ and a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values. We further assume a set of actions $\mathcal{A} = \{A_1, \ldots, A_m\}$ and a utility function $u : \mathcal{A} \times \mathcal{V} \to \mathbb{R}$. To select an action, take R samples, $H^{(1)}, \ldots, H^{(R)}$, using $\mathcal{M}$, and let best be the set of actions that receive the largest summed utilities, i.e.,
>
> $$\textsc{best} = \{A_j : \sum_{i=1}^{R} u(A_j, H^{(i)}) \text{ is maximal}\}.$$
>
> Take action $A_j \in \textsc{best}$ with probability $\frac{1}{|\textsc{best}|}$.

Again by the Law of Large Numbers, Rule B will approximately maximize the expectation of u (recall §2.7 above). Decision Rule B is applicable in case our "intuitive model" $\mathcal{M}$ generates states H in $\mathcal{V}$, and we have some way of looking up $u(A, H)$ for each of the relevant actions A in $\mathcal{A}$. Such a rule would be appropriate insofar as probabilistic information, which we assume is encoded by generative models, and utility are distinctly represented (see, e.g., Dayan and Daw 2008).

Compare Decision Rule B with Decision Rule C, which, so to speak, folds the actions and utilities right into the generative model $\mathcal{M}$. Instead of generating states, we can imagine $\mathcal{M}$ generating action-state-utility triples, thereby defining a joint distribution over such triples.

> Decision Rule C: Everything is as in Decision Rule B, with possible states $\mathcal{V}$, actions $\mathcal{A}$, and utility function u. The only difference is that our

> model $\mathcal{M}$ directly generates state-action-utility triples $\langle H, A, U\rangle$, and thus conditioning on action $A_j \in \mathcal{A}$ generates state-utility pairs $\langle H, U\rangle$. Taking R samples from $\mathcal{M}$ conditioned on A_j produces $\langle H^{(1)}, U^{(1)}\rangle, \ldots, \langle H^{(R)}, U^{(R)}\rangle$. Let $\textsc{sum}_j = \sum_{i=1}^{R} U^{(i)}$ and $\textsc{best} = \{A_j : \textsc{sum}_j \geq \textsc{sum}_k \text{ for all } k \leq m\}$. Take $A_j \in \textsc{best}$ with probability $\frac{1}{|\textsc{best}|}$.

Both rules B and C generalize Decision Rule A from §2.6, and they can be used interchangeably. However, Decision Rule C is applicable when the model $\mathcal{M}$ itself incorporates actions and utilities, and states are viewed not as mappings from actions to utilities, but as *outcomes* of actions. The intuition—at least in the case of high-level reasoning and planning—is that we are simulating different courses of action leading to different outcomes, and the utilities are somehow *extracted* from how we react to those outcomes. For instance, one can imagine utility being derived from affective response or some other aspect of the resulting mental state.

There are a number of intriguing variations on Decision Rule C, given this more expansive flavor of generative model. For instance, we could imagine a Decision Rule D that does not condition on each action in turn, but rather generates state-action-utility samples and takes the action that has best sample-average utility. Thus, stochastically choosing which action to simulate would be part of the generative process. This might be particularly attractive from the perspective of time and resource-boundedness, since one might only be able to consider a few potential actions. An effective use of this generative process would be to test the actions roughly in order of their *a priori* chance of being considered optimal (cf. Chapter 4).

An even more radical departure from rules B and C (call it E) would be to condition $\mathcal{M}$ on some statement about utility, e.g., that it is maximal, in order to infer an action directly. This is essentially the idea behind the *planning as inference* paradigm (Solway and Botvinick, 2012), which has been shown to account for a number of central phenomena associated with goal-directed behavior, and has been related to specific neuroanatomic structures. This program is quite complementary to the Sampling Hypothesis (Solway and Botvinick, 2012, 142), in that the actions themselves could be sampled, to be used by a rule much like Decision Rule A. In this case, the generative model is interpreted as directly defining the probability that a given action is best in the current context.

Once again, all of these variations—B through E—essentially extend Decision Rule A to the setting in which $\mathcal{A} \neq \mathcal{V}$ and the utilities are possibly more complex. Which of these is appropriate will depend on what the brain is tracking, and the answer may obviously depend on the cognitive function. Perhaps sometimes we are tracking states of the world independent of our own actions, while other times we are directly tracking the probability that a given action is the one to take. (Once again, notice that the Yang and Shadlen 2007 results are compatible with either interpretation.) There is a great deal of exciting work exploring these questions, though the role of sampling in more complex decision problems has yet to be explored in depth.[16] Happily, the Sampling Hypothesis by itself is neutral on this issue, and the evidence reported in §2.8-§2.10 assumes nothing more specific than something along the lines of Decision Rule A.

2.12 Measuring Subjective Probability

As mentioned already several times, there has been much skepticism about what one can conclude about subjective probabilities from the traditional methods of elicitation, viz. assessing fair betting odds, preferences between gambles, etc., which we inherit from Ramsey, de Finetti, Savage, and others. If people's choice behavior is grounded in internal stochastic sampling, this casts yet further doubt on the possibility that these methods will reveal psychologically real probability representations. Just as with any stochastic system, sometimes we do not learn enough from a single observation. The variability of individual choice behavior has been recognized for a long time, and alternative methods have been developed for measuring and predicting choice behavior (Thurstone, 1927; Luce, 1959; Luce and Suppes, 1965). Unlike in much of this work—which, like Ramsey, de Finetti, etc., had a tendency toward behaviorism—the current goal is to understand and measure internal states and computations, with choice and behavior merely one source of evidence.

[16] An important exception is work on the *Decision by Sampling* framework (Stewart et al., 2006), which is compatible with, but orthogonal to, the Sampling Hypothesis as understood in this chapter.

The empirical work described in the previous three sections represents diverse approaches to investigating subjective probabilities, provisionally conceived as internal sampling propensities. This includes comparing models to what we know about neural dynamics and basic psychological functions (§2.8), investigating links between characteristic biases produced by humans and by sampling algorithms over hypothesized distributions (§2.9), and paying close attention to population-wide and individual variation in behavior (§2.10). Much of this work comes from the "rational analysis" tradition of Bayesian psychology—to be discussed in more depth in the next chapter—which typically derives a probabilistic analysis from first principles, or from the true statistics of the environment, and then compares behavior to the ideal model. Further sophisticated methods have been developed very recently for discovering subjective probabilities directly, including sequential experimental techniques that adaptively depend on subjects' responses, in effect using subjects themselves to define a Markov chain converging to some distribution.[17]

In view of the discussion of utility in the previous section, this plurality of methods I take to be an obvious improvement over the traditional operationalization of subjective probabilities by means of betting behavior, at least if we are interested in discovering concrete psychological mechanisms. Together, the results from §2.8-§2.10 make a strong case that the Sampling Hypothesis is onto something. However, there remains one criticism of the use of probability from §2.3 that we have not yet addressed, namely the charge of arbitrary precision. The models we have considered, such as the Boltzmann Machine, involve arbitrarily precise real number values. What justifies using precise values to characterize subjects' mental states?

Recall the two main sources of criticism from §2.3.2: (1) that introspection does not seem to ratify real number values, and (2) that measurement of subjective states cannot distinguish between arbitrarily close values. I take it that the response to (1) is now clear. Subjective probability, on the sampling view, is not the sort of mental state to which subjects have privileged or immediate access (recall Footnote 7). Criticism (2) is perhaps more pressing.

[17] See Lewandowsky et al. (2009) and Sanborn et al. (2010) for two different approaches. A full discussion of this very intriguing work is beyond the scope of this chapter.

If the Sampling Hypothesis is correct and subjective probabilities can be thought of as sampling propensities related to concrete generative processes, then the question of how to characterize the mental states of subjects is in fact a special case of the more general question of how to characterize partially observable stochastic processes. As with any indirect measurement, we must first make certain assumptions about the structure of the domain in question. Provided we can agree about some of these general structural assumptions—e.g., what kind of generative process it is, what the set $\mathcal{V}$ of possibilities looks like, etc.—one of the central remaining issues is how to determine the probabilities associated with the generative process. The question of whether to use precise real numbers in these estimations, or something like intervals or sets of probabilities, is a matter of general methodological principle.

The theorist can make further assumptions about the underlying distribution and, e.g., estimate parameters of the assumed distribution from data through Bayesian analysis (Kruschke, 2010), or perhaps make a specific assumption about the exact distribution and make an inference about the decision rule, such as the number R of samples subjects are drawing (Vul et al., 2013), or the number of particles in a particle filter (Levy et al., 2009; Sanborn et al., 2010). Other theorists may be reluctant to make any further assumptions and prefer to use the data more directly to establish intervals for the relevant quantities, or to use more traditional statistical estimation techniques. This of course involves us in fundamental methodological issues that go beyond the scope of this thesis. The important point to make here is that, once we have accepted the sampling view as a working hypothesis, further questions about how to estimate specific subjective probabilities—including whether to use probabilities, intervals, etc.—are no different from questions about how to model any other uncertain phenomenon. There is nothing special about the fact that it is mental states we are trying to measure. I take this as one of the benefits of identifying subjective probability with a kind of concrete propensity.

2.13 Conclusion and Preview

According to the Sampling Hypothesis, subjective probabilities are identified, not directly with propensities to act as Ramsey suggested, but with sampling propensities of internal generative models. Like the more traditional analyses of subjective probability, which sometimes eschew as evidence subjects' explicit reports about what they believe, the sampling view is grounded in concrete, measurable phenomena. The hypothesis is distinctive in that the target phenomena are stochastic processes associated with internal generative "intuitive models", which are assumed to mirror the probabilistic dynamics of what they are modeling. The evidence reviewed recommends this view, at least for a wide variety of important cases of prediction and judgment, as well as more automatic "subpersonal" psychological processes. I have tried to show that the Sampling View allows us to be realists about subjective probability, in such a way that these probabilistic representations play a genuine and central role in the production of behavior, while avoiding some of the most common objections to the idea that probabilities are encoded in the mind.

There are many large open questions and issues raised by the resulting viewpoint; here I would like to mention just two, as links to the next two chapters.

The first question is how we can make progress in understanding the nature of the models we are assumed to be sampling from. The Bayesian program in psychology offers one promising strategy for making headway on this question, by mapping out a "top-down, problem-first" methodology for understanding human psychology. Conversely, from the point of view of the Bayesian program, the Sampling View suggests responses to a number of challenges that have been raised for the program, especially in recent years (some overlapping with those reviewed above in §2.3, but some quite distinct). The topic of Bayesian models has surfaced at various points in this chapter, and it will be the main topic of the next chapter.

The second question is perhaps more directly philosophical and concerns how we should think about rationality on the resulting picture. In particular, the sampling view opens up the possibility of a novel and substantive study of *bounded* or *procedural rationality* (Simon, 1955, 1956, 1976). Instead of maximizing expected utility with respect to the limiting behavior of one's internal generative model, it may be significantly more cost-effective to draw only a few samples and make a decision on that basis. Understanding this kind of rationality is important for the "rational analysis" research strategy associated with the Bayesian program, so it will be central in our discussion in the next chapter. But it also opens up fascinating lines of inquiry related to *metareasoning* and the idea that it sometimes pays to reason about one's own reasoning. This theme has surfaced several times in this chapter as well. It comes up both in questions about how to use our limited resources in an optimal way, and, relatedly, with the idea that observations about our own reasoning processes can be treated as evidence relevant to a decision problem. One of the benefits of taking a more algorithmic view of the mind is that we can begin to understand how and why metareasoning can be useful. At the same, as I will try to show, it introduces complications and subtleties into both psychological theorizing and our understanding of what it means to be a rational agent.

Chapter 3

Boundedly Rational Analysis

3.1 Introduction

In the previous chapter, we encountered a number of strategies for discovering aspects of cognitive processing, including "bottom-up" methods based on careful studies at the neural level, as well as "top-down" methods that begin by positing an underlying problem to be solved and comparing experimental data to the ideal solution to this problem. For low-level tasks, e.g., aspects of perceptual processing, there is beginning to emerge an impressive confluence of these two strategies (Knill and Pouget, 2004; Yang and Shadlen, 2007). However, for higher-level cognitive functions, it remains the case that what we know about how the brain works severely underconstrains the possible algorithms capable of reproducing the input-output patterns observed in human behavior. What is needed is some way of further reducing this space of possibilities, and ideally of suggesting new experiments that would allow reducing it even further. Partial suggestions of this nature have been made through the decades. There is an old tradition of using reaction-time data to distinguish different hypotheses about information processing (a classic example in this vein is Sternberg 1969 on memory retrieval), which can also be used to guide hypothesis formation. Researchers in the connectionist tradition have taken observations from what we know about how brains are spatially organized as starting points for homing in on specific models (Rumelhart et al. 1986), and have often brought lesion studies, and other related methods, to bear on specific hypotheses.

A different, but compatible, approach, epitomizing the "top-down" method of investigation, is the idea of *rational analysis*, pioneered by Marr and Poggio (1976); Marr (1982), and greatly extended by Anderson (1990) to include the study of many familiar high-level cognitive tasks. The basic idea is to try to understand a given cognitive function as the mind's solution to some underlying problem. By formalizing the underlying problem sufficiently, one can derive an ideal, or at least reasonable, solution to the problem. A crucial step in rational analysis is to devise appropriate experiments so as to compare human performance with this purportedly ideal solution. To the extent that performance diverges from the ideal, that may suggest ways the original formulation of the problem was mistaken; one can then attempt to reconsider the problem formulation, repeating these steps, and iterating with the hope of zeroing in on a better understanding of how the brain is carrying out a task.

The rational analysis program is quite general and can be applied to virtually any "black-box" problem, in which the inner workings of some system are either unobservable or too complex to understand with our current tools. For the specific kinds of tasks that the human mind typically faces, it is common to formulate them as involving inference problems under uncertainty (Oaksford and Chater, 2007), or perhaps more generally, decision problems under uncertainty (Anderson, 1990; Maloney and Mamassian, 2009), for which Bayesian methods are widely regarded as providing a robust and well-understood notion of optimality or rationality (Bernardo and Smith, 1994; DeGroot, 2004). It should then perhaps be no surprise that the recent wave of activity in the so-called Bayesian program in psychology over the last decade and a half identifies itself as carrying out the rational analysis program.

Despite its impressive success in modeling a wide variety of cognitive phenomena—arguably unprecedented in range and scope—the Bayesian program has met considerable criticism over the past several years, from both psychologists and philosophers (Gigerenzer and Goldstein, 1996; Kwisthout et al., 2008; Eberhardt and Danks, 2011; Jones and Love, 2011; Bowers and Davis, 2012; Marcus and Davis, 2013). Some of the criticisms are quite general to the rational analysis strategy, and mirror some of the same criticisms of the adaptationist program in evolutionary biology (cf. the well known criticisms by Gould and Lewontin 1979). But others are specific to the Bayesian incarnation of the program in psychology. In this chapter I would like to focus on four of the main criticisms in this vein that have surfaced in this literature.

(1) The computations that seem to be required by many Bayesian models are intractable. This makes their application to human cognition questionable at best (Gigerenzer and Goldstein, 1996; Kwisthout et al., 2008; Jones and Love, 2011).

(2) There is a general problem of how a Bayesian model is confirmed by the data. Across many experimental studies, subjects show a kind of "posterior matching" instead of what would seem to be the ideally rational strategy of making the maximum *a posteriori* (MAP) guess (Eberhardt and Danks 2011; Marcus and Davis 2013, cf. §2.10). On the face of it, this rather disconfirms the hypothesis that subjects are optimal or rational.

Our discussion in the previous chapter suggests a possible answer to these first two criticisms. If subjects are sampling from the appropriate distributions, then computational difficulties are at least mitigated, and we can account for matching behavior. What is less obvious is the sense in which the *Sampling Hypothesis* follows from, or is even consistent with, a rational analysis. In particular:

(3) If people are merely approximating Bayesian solutions, then in what sense is their behavior Bayesian, or rational at all (Eberhardt and Danks, 2011)? Moreover, once we give up on perfect Bayesian rationality, why should we be constrained to search for approximations to the Bayesian solution in the first place? Might not other solutions be better for the task at hand (Gigerenzer and Goldstein, 1996; Kwisthout et al., 2008; Eberhardt and Danks, 2011)?

While I believe there is considerable force to criticism (3), I also think it is not insurmountable, provided we understand the "rational" in rational analysis as "boundedly rational" in the right way. This suggestion has been made in the Bayesian literature, though sometimes in a way that presupposes, rather than justifies, a Bayesian analysis. One of my aims in this chapter is to outline what it would take to derive, or at least be justified in hypothesizing, an approximately Bayesian analysis of some phenomenon, without merely presupposing that the analysis should be based on such an approximation. As I will try to explain, answering this question satisfactorily will require simultaneously tackling an equally difficult problem:

(4) How are we to determine when a given problem or cognitive function can be appropriately modeled as (approximate) Bayesian inference? Conversely, when should we conclude that a given phenomenon should not be understood as Bayesian inference (Marcus and Davis, 2013)?

Whereas certain low-level perceptual processes are widely held to carry out Bayesian calculations, there are other cases where it would be outlandish to propose a Bayesian rational analysis, e.g., in understanding people's naïve judgments about particle physics, or how best to run a company. Clearly, a Bayesian rational analysis works well in some cases, less well in others. As Anderson (1991b) put it, rational analysis is a "high-risk, high-gain enterprise" (472): high-gain because we stand to achieve a better understanding of how the mind works and in what ways it is optimized, but high-risk because in many cases a rational analysis is simply inappropriate to the task under investigation. One might like at least some help in determining when we should expect a Bayesian rational analysis to provide insight and when not.

While the reader should not expect easy, clean answers to these questions, I do aim to make some progress in what follows. My claim will be that we can start to answer these questions by shifting the focus from perfect Bayesian (inferential) rationality to something closer to *bounded rationality* (Simon, 1955). Very simply, the extent to which we should expect an approximately Bayesian (e.g., sampling-based) analysis of some cognitive capacity to be adequate depends on whether the method in question provides a boundedly rational solution to the problem of maximizing some end, subject to computational and other constraints. To be sure, demonstrating such rationality in any interesting case is a daunting task, so part of the burden will be to find broad characteristics of good solutions to problems faced by cognition, so that we can narrow down the search space before making more precise comparisons between competing boundedly rational algorithms.

After outlining the rational analysis program as put forth by Anderson (1990, 1991b), and giving an example of an uncontroversial example (among many) where rational analysis has led to important progress (§3.2, from Movellan and McClelland 2001), I will then give an example of a more substantive rational analysis with work by Sanborn et al. (2010), building on Anderson (1990, 1991a), which involves the crucial step from an ideal "computational level" model to a (sampling-based) approximation, to explain features of the behavioral data (§3.3). For reasons to be discussed, I take this also as a successful case of rational analysis, showing how criticisms (1) and (2) can be met in a concrete example. Using it as an illustration, in §3.4 I will revisit and expand upon challenges (3) and (4), and explain why considering bounded, or procedural, rationality helps answer to these challenges. In §3.5, I develop some tools for understanding boundedly rational agents, the main application of which will be to *retrodict* that the model offered by Sanborn et al. (2010) ought to provide a better fit to the human data than the original model of Anderson (1990, 1991a). Through computational simulation, I will explore the respective fitness of these and other agent models. I will also present examples of agents in environments for whom a Bayesian analysis is inappropriate, for reasons concerning bounded rationality, which I will argue partly responds to challenge (4). Finally, as a link to the next chapter, in §3.7 I will introduce complications to the picture when we consider agents capable of reasoning about their own reasoning, for whom the question of which computations to perform in the course of solving a problem itself becomes part of the problem to be solved. Again through computational simulation, I will show that agents who spend time and resources making "metalevel" decisions can be more boundedly rational than those who do not.[1]

3.2 The Rational Analysis Program

Rational analysis is premised on the assumption that a given cognitive function or system is, in a sense, optimized to carry out some task in a given environment. This

[1] A short version of this chapter has since appeared in the Proceedings of the 36th Annual Cognitive Science Society Meeting (Icard, 2014). See also Griffiths et al. (2014), which has since appeared and presents related, though in some important ways different, views.

idea by itself is not new, and can be traced back to Brunswick, Dewey, and probably further. The idea of rationalizing an agent (or artifact) in terms of purported beliefs and goals, for the purpose of explaining and predicting behavior, is familiar from Dennett's *intentional stance* in philosophy of mind (see, e.g., Dennett 1981b).

Rational analysis in psychology proposes that this working assumption can support a general methodology for discovering how cognitive systems work, not just for the purpose of prediction. By assuming optimality or rationality, correctly specifying the problem *ipso facto* narrows down the cognitive mechanism: it must be one that rationally solves the problem. To the extent that this assumption is confirmed and borne out by carefully devised experimentation, this in turn provides a *rationalization* of the behavior. We then have what is sometimes called a "computational-level" explanation, which is valuable even in cases where we know how the system works at a fine level of grain, that is, even when we have an "algorithmic" or "implementation-level" understanding (Marr, 1982; Pylyshyn, 1984). It gives us a sense of what the given system is doing at a high level (Marr and Poggio, 1976).

While in the background is an assumption that agents (or parts of agents) have in some way *adapted* to satisfy their goals in their typical environments, proponents have typically remained neutral on the question of whether this is at an ontogenetic or phylogenetic level, or both. In order for the program to be truly explanatory, it is important to make the nature of that assumption explicit (Danks, 2008). But for the purpose of the methodological strategy, which is our focus in this chapter, it makes little difference. For this purpose, rationality or optimality is less of an *a priori* assumption than a guiding principle. Its having worked in the past gives us some reason to hope it will work in analogous future circumstances. Of course, one of the challenges will be to determine the sense in which it has worked in past applications—that is, that behavior claimed to be rational is indeed rational in a clear sense—which is part of what has been called into question (challenge (3) above).

Another important assumption on which rational analysis relies is a kind of *modularity*. Full-blown modularity in the sense of Fodor (1983) is not necessary, but modularity of *function* is critical. We need to be able to identify separate cognitive processes by the functions they carry out, and by the goals that purportedly guide them (cf. Pylyshyn 1984). Obviously, in many cases an agent's goals will be subserved jointly by many cognitive functions working together. Understanding an utterance, for example, will involve audial perception, syntactic parsing, memory retrieval, categorization, social reasoning, and much more. The possibility of performing a rational analysis of each of these processes individually depends on our being able to factor the overall problem (e.g., understanding an utterance) into separate subproblems (identifying an appropriate syntactic structure, etc.). Even when this factorization is conceptually possible, one of the main criticisms—famously made against the adaptationist program in biology by Gould and Lewontin (1979)—is that such modules or parts of agents are rarely what is optimized or improved through evolution. I will not try to explicate or defend the assumption here, but it is an important topic that deserves further attention; and it will resurface below in §3.7.2.

3.2.1 Example: The Morton-Massaro Law

Before moving on to more substantive versions of the rational analysis strategy, it may be helpful to begin by illustrating the main idea with a more modest application.

In the study of perceptual inference—specifically the integration of multiple sources of perceptual information—there was a question about whether these multiple sources are processed separately and then integrated to produce a response (the *Morton-Massaro model*), or whether the different sources may affect the contribution each makes to producing a response, giving rise to context effects (the *interactive activation hypothesis*). While in principle one might try to distinguish these two hypotheses simply by looking inside the brain and seeing whether perceptual information flows in only one direction, this strategy is obviously prohibited by our limited understanding of the relevant anatomical mechanisms. Faced with this difficulty, Movellan and McClelland (2001) sought to understand the situation better by pursuing a rational analysis of perceptual integration.

We can think of the problem of perceptual integration as that of inferring some latent state of the world H from perceptual cues $D_1, \dots, D_n$. If we think of the latent state as causing the cues, and the agent is assumed to have some prior expectation about how probable each state is, we can determine the relevant probabilities using Bayes Rule:

$$P(H \mid D_1, \ldots, D_n) \;\propto\; P(H)P(D_1, \ldots, D_n \mid H)\ .$$

The assumption is that this specifies the goals of the perceptual system, in that the system is well adapted, or optimized, to the extent that it has a good model of the world.[2] Movellan and McClelland then derived what would have to be the case about the agent's environment in order for the Morton-Massaro model to be optimal. Simplifying slightly, they found that it would be optimal only if the so-called naïve Bayes assumption is valid:

$$P(D_1, \ldots, D_n \mid H) \;=\; \prod_{i \leq n} P(D_i \mid H)\ .$$

This holds when the likelihood term in Bayes Rule can be factorized so that each cue variable is conditionally independent of all the other cue variables. This in turn suggested new experiments to test whether the Morton-Massaro accounted for the human data even when the naïve Bayes assumption does not hold. One such context is in perceiving words (H) given noisy data about the marks making up letters ($D_1, \ldots, D_n$). Movellan and McClelland found that people are indeed sensitive to context effects in these tasks, in a way that could not be explained by the Morton-Massaro model. On the other hand, in many other tasks where the naïve Bayes assumption holds, e.g., in the problem of integrating audial and visual speech information, the Morton-Massaro model is quite accurate in its predictions. Thus, a proper rational analysis shows precisely when we should expect the Morton-Massaro generalization to hold, and just as importantly, suggests that an integration *mechanism* should at least be capable of capturing the non-factorizable cases, favoring something more like the interactive activation hypothesis.

This is a rather modest example of rational analysis, in the sense that the class of possible mechanisms under discussion is already narrowed down considerably. But it

[2] For brevity, I am ignoring the fact that what really matters is the perceptual response, which is presumably some function of H. Costs are also assumed to be negligible in this example. Both of these will take center stage later in §3.5.

does show how the strategy, when applicable, can be used to suggest new experiments and guide inquiry. It also clarifies how understanding the environment and the task at hand can explain aspects of our cognitive processing. In this case, context dependence in perception is explained by statistical features of the environment combined with the evident goal underlying the perceptual system.

3.2.2 Anderson's Six Steps

A more substantive use of the rational analysis strategy is to help *derive* a cognitive model from first principles, using rationality as a guide. Rather than merely adjudicating between individual competing models, the idea is to *discover* possible models by studying what a reasonable or rational solution to some underlying problem would be, given information about computational limitations and the statistics of the environment. This more controversial version of rational analysis was pioneered in cognitive psychology by Anderson, who codified the methodology in six steps (Anderson, 1990, 29):

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

Anderson gave illustrations of the methodology in four domains, all well-studied in cognitive psychology: memory retrieval, category learning, judging causal strength, and problem solving (i.e., decision making). In each case, he was able to show an impressive fit to a wide range of data, demonstrating that the strategy can lead to interesting results. In the next section, we illustrate this method using the example of category learning, beginning with Anderson's own analysis, then turning to a recent development by Sanborn et al. (2010).

3.3 Rational Analysis of Category Learning

Anderson (1990, 1991a) characterized the problem of category learning in terms of prediction. A subject observes a sequence of objects with various combinations of features; upon observing a new object, the subject needs to make an inference about some unobserved feature. Let Y_i be the value of the feature of interest for the ith observed object, and let X_i be the ith vector of values for the remaining features. For instance X_i could consist of some visual properties, while Y_i corresponds to a label or other property of interest (e.g., being dangerous, edible, etc.). Let us furthermore abbreviate the conjunction of the first n values, $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_n$, as $\mathbf{Y}_N$ and $\mathbf{X}_N$, respectively. The problem facing the subject is thus to infer the value of Y_n, given observations of $\mathbf{X}_N$ and $\mathbf{Y}_{N-1}$, i.e., to find the value of Y_n that maximizes the posterior probability $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ of Y_n given $\mathbf{X}_N$ and $\mathbf{Y}_{N-1}$.

Anderson argued that the ideal way to determine such probabilities would be to consider all possible clusterings, or partitions, of the first n objects, figuring out the probability of each clustering, and using the clusterings to determine the probability of Y_n given X_n. In other words, the real task of category learning is to determine the value of a latent variable Z_n, which corresponds to a clustering of the n objects observed so far. Then we can determine the posterior probabilities by summing over the possible clusterings:

$$P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;=\; \sum_{Z_n} P(Y_n \mid Z_n)\; P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})\;. \tag{3.1}$$

Anderson defined these terms based on some casual assumptions about a typical environment, e.g., consisting of living things or artifacts. In defining $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$, he invoked Bayes Rule:

$$P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;\propto\; P(\mathbf{X}_N, \mathbf{Y}_{N-1} \mid Z_n)\; P(Z_n)\;. \tag{3.2}$$

The first term is just the likelihood of the features given a clustering, which we also need to compute $P(Y_n \mid Z_n)$ in Eq. (3.1), and is given by a beta distribution. For large n this approaches the relative frequency of similarly categorized objects that had these features. As Sanborn et al. (2010) showed, Anderson's prior term $P(Z_n)$ is equivalent to a *Dirichlet process mixture model*, which implements a "rich get richer" scheme, meaning that new objects are more likely to fall under categories that already have many members. The technical details involved in these computations can be found in the Appendix. They are not crucial at this point.

The important observation in the present context is that the computations required by Eq. (3.1) are wildly intractable. As Anderson already pointed out, the number of clusterings Z_n as a function of n grows exponentially. For $n = 10$, there are already $115,975$ possible clusterings, making the sum in Eq. (3.1) intractable in all but the simplest of cases. This is not an atypical feature of "ideally rational" Bayesian models either, and has been a significant point of criticism, as discussed earlier in §2.3.1 and §3.1. It is exactly at this point where one might wonder what we should expect the "ideal" Bayesian analysis, as codified in Eq. (3.1), to tell us about human categorization.

Of course, Anderson was well aware of this, which is why he included step 3 in his methodology. In addition to the obvious constraint that the required computations should be tractable, Anderson also assumed that at any given time, a subject ought to have settled on a particular clustering of objects seen so far, so that as new objects are observed, the only question is how to extend that partition to include the new object. This led him to the following proposal:

> Anderson's Local MAP Algorithm: Upon observation of a new object with features X_n, let Z_n^* be the extension of the current partition Z_{n-1} that maximizes $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$. One can then estimate $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ by merely calculating:
>
> $$\tilde{P}(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \; = \; P(Y_n \mid Z_n^*) \, P(Z_n^* \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \, . \tag{3.3}$$

That is, instead of summing over all partitions every time one needs to make a prediction, Anderson's local algorithm has the subject deterministically choosing the maximum *a posteriori* (MAP) partition following each new data point. Eq. (3.3) is supposed to be a tractable version of Eq. (3.1).

Impressively, Anderson showed that his local MAP algorithm was able to account for a wide array of empirical phenomena collected from over two decades of work on categorization, including order effects, prototype effects, the relative ease of learning different categories of Boolean concepts, and many more (see Anderson 1990, 1991a for discussion).

Despite this success, from the point of view of rational analysis, the move from Eq. (3.1) to Eq. (3.3) may seem somewhat arbitrary. In what sense is the local algorithm optimal or rational once we take tractability into account? One can imagine any number of ways of avoiding having to sum over all partitions. Why must we always settle on a single partition after each new observation? And why ought we do this deterministically?

It is here that Sanborn et al. (2010) pick up the line Anderson started. Instead of considering a deterministic simplification of the problem in which the subject chooses the MAP clustering at each point, they introduce the *particle filter*, a sequential Monte Carlo sampling algorithm, as a possible algorithmic-level model. We discussed this algorithm briefly in §2.8. In the present setting, the idea is that a subject maintains at any given time a set of R "particles", each corresponding to a clustering, and bases inferences on the whole set:

> Particle Filter Algorithm (Sanborn et al., 2010): Upon observation of a new object with features X_n, draw samples $Z_n^{(1)}, \dots, Z_n^{(R)}$ from $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$. One can then approximate $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ by calculating:
>
> $$\tilde{P}(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;=\; \sum_{r=1}^{R} P(Y_n \mid Z_n^{(r)})\; P(Z_n^{(r)} \mid \mathbf{X}_N, \mathbf{Y}_{N-1})\;. \qquad (3.4)$$

Anderson's MAP algorithm would be a special case of the particle filter algorithm with a single particle ($R = 1$), except that the particle is drawn deterministically as the MAP clustering. Unlike the MAP algorithm, whose response according to Eq. (3.3) is a deterministic function of the input, the particle filter algorithm's Eq. (3.4) converges to Eq. (3.1) in a certain sense: as we draw R samples from $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ multiple times, the estimated distribution of responses asymptotically approaches the true distribution in Eq. (3.2). Furthermore, the expected distance of the approximate distribution from that given by Eq. (3.2) reduces with the number R of particles. This allowed Sanborn et al. to estimate the true distribution from Anderson's model (as given by Eq. (3.1)) very closely, with 100 particles, and compare the resulting model to the human data. While the fit was good, it was not as good as Anderson's MAP algorithm, which better mirrored characteristic biases and memory limitations on the part of human categorizers.

Intriguingly, Sanborn et al. were able to show that the particle filter algorithm with only one particle was able to account for all of the same empirical phenomena that Anderson had shown for the MAP algorithm, and in fact it provided an even better fit to some of the data. In general, the MAP algorithm does not allow for enough individual variation, which is exhibited clearly across empirical studies. For example, consider a case where two possible partitions Z and Z' have nearly equal probability given the data seen so far. If the probability of Z is only slightly higher than that of Z', then the MAP algorithm will always use Z to determine later values of Y, whereas the particle filter algorithm will "split the difference", using Z only slightly more than half the time. The latter pattern is much more in line with the pattern observed in human data. Given our discussion of this issue (§2.10, §A.2), we might conclude that this population-level "posterior matching" behavior should be reflected in the stochastic nature of each individual's decision procedure. The analysis by Sanborn et al. (2010) suggest that this is a reasonable hypothesis in this case.

3.4 Toward Boundedly Rational Analysis

The work described above by Anderson (1990) and Sanborn et al. (2010) points the way toward a response to criticisms (1) and (2) from §3.1. Both the MAP algorithm and the particle filter algorithm are tractable versions of the "ideal" inference, and both fit the human data quite well, with the particle filter fit potentially better. Moreover, both are at least vaguely inspired by the computational level model described by Eq. (3.1), even though neither is strictly *derived* from that equation. The particle filter algorithm is, in a precise sense, an approximation to the intractable Bayesian computation that would be the "optimal" solution to the problem, if an agent had free, unbounded computational resources.

Now having concrete examples on the table, I would like to raise three principal questions, to sharpen challenges (3) and (4) from the introduction:

I. Having made a computational level analysis and observing that it involves intractable computations, when should we expect the computational level analysis nonetheless to constrain the class of possible algorithms?

II. What is the relation between the "problem being solved" when we ignore computational limitations, and the "problem being solved" when we take such considerations into account?

III. In what sense might an approximation that asymptotically converges to an "ideal" solution be *more rational than* another algorithm that does not?

An uncharitable reading of some of the Bayesian literature might have one think the assumed answers to I, II, and III are "always", "identity", and "patently", respectively. There have been several recent papers investigating *resource rational* models of cognition, showing that one can derive optimal sampling behavior by trading off cost of sampling with improvement from better approximation (Vul et al., 2009, 2013; Lieder et al., 2012). Yet, these analyses, discussed further in §3.6 below, do not by themselves offer convincing answers to I-III, because they assume from the start that the agent makes decisions by sampling from an internal (implicit) distribution. One large and important question is why it would make sense to represent such a distribution in the first place, and why it would make sense to sample.

As critics have pointed out (Kwisthout et al., 2008; Eberhardt and Danks, 2011), the behavior predicted for an agent that uses an approximation algorithm is different from that predicted for an agent following exact Bayesian reasoning. Indeed, this difference turns out to be crucial for explaining the data. As discussed above, the two approximation algorithms account for a number of patterns in human behavior that are inconsistent with the "ideal" model, such as order effects. In this particular case, it might seem as though Anderson and Sanborn et al. were lucky, in that their constrained solutions to the problem, both conceived as simplifications of the ideal model, happened to provide a reasonable fit. This is a "top-down" strategy *par excellence*, but at this point there seems to be an element of arbitrariness. Before looking at the data and seeing whether some version of one of these models would match it, how could we have known that this was a promising strategy? Both Anderson and Sanborn et al. pitched their alternative algorithms as "rational approximations to rational models" and thus as following from the rational analysis strategy. But insofar as they predict behavior that diverges from the purportedly optimal behavior, we cannot simply conclude from their being approximations that they are optimal or rational in any important sense. It may be that once we give up on optimality, some totally different, perhaps simpler, algorithm will be more rational than, or otherwise preferable to, an approximation derived from the computational level analysis. We do not yet know how to judge this in advance.

Thus we are led to consider question I above: when should we expect the computational level analysis to suggest a reasonable algorithmic level analysis? To answer this question properly, I believe it is important to consider a different answer to question II. The "problem being solved" at the level of concrete, implementable algorithms is not the same as that considered at the computational level, in general. Whereas the computational level problem or goal can often be characterized as a Bayesian *inference problem*—e.g., predict the correct category given all the data—the real problem being solved is more like a *constraint optimization problem*: make the best guess subject to memory, time, energy, and other cost constraints.

This general claim is of course quite old. That the normative standard against which to measure human behavior is one of *bounded rationality* goes back to Herbert Simon (1955; 1956). Simon made the crucial distinction between *substantive rationality*, which is characteristic of computational level accounts, and *procedural rationality*, which concerns how concrete, bounded agents manage their limited resources to solve problems (Simon, 1976). All we should expect of human agents, according to Simon, is that they are boundedly or procedurally rational. Many following Simon have demonstrated concrete cases where people do seem to use simple strategies that could not have been easily derived from a computational level analysis. Occasionally, it is even possible to show that these strategies are better adapted to the task environment than what one might have derived from a top-down rational analysis, e.g., as in Gigerenzer and Goldstein's (1996) analysis of their "Take the Best" algorithm (though see Lee and Cummins 2004, which suggests the matter is more subtle).

Taking the perspective of bounded rationality may seem to be at odds with the rational analysis strategy, at least as conceived by Anderson and by the Bayesian tradition following him. Recall that step 3 from Anderson (1990) required making only the *minimal* assumptions about computational limitations. Some authors have been quite explicit that the force of rational analysis is lost by making too many assumptions about cost, mechanism, or other algorithmic level issues (Anderson, 1991b; Daw et al., 2008). From the other side, Gigerenzer and colleagues have been equally adamant that there are fundamental differences between the "fast and frugal heuristics" viewpoint and the Bayesian viewpoint. Herbert Simon himself in fact commented on Anderson (1990), also arguing that rational analysis would be anathema to the idea behind bounded rationality:

> Bounded rationality is what cognitive psychology is all about. And the study of bounded rationality is not the study of optimization in relation to task environments. (Simon, 1991, 35)

A given problem, according to Simon, could be solved reasonably efficiently by many alternative strategies, and "the environment cannot predict which of these alternatives will govern the adaptive behavior" (Simon, 1991, 35). For the case of categorization in particular, Simon anticipated our current predicament, noting that many possible algorithms are conceivable:

> Our interest is in the learning process itself – not a hypothetical one or an optimal one, but the one that people use. We want a learning theory precisely because people do not arrive at optimal classifications immediately or costlessly. (35)

Simon was skeptical that optimality considerations would ever lend such insights.

I believe there is more than a grain of truth coming from both sides. Simon and followers are right that what really matters for understanding and assessing human (and other) agents is something more like procedural, bounded rationality. But from the other side, I believe that there are many clear cases where Bayesian analyses have in fact led to insights into the nature of psychological processes. The example from Movellan and McClelland (2001) illustrates this nicely, in a case where the hypothetical mechanisms are already narrowed down, and the computational problem is reasonably tractable. For more difficult problems the situation is less clear *a priori*. I agree with Gigerenzer et al. that we should not always assume we can simply choose from among approximations to computational-level models. On the other hand, as Anderson (1990, 1991a), Sanborn et al. (2010), and others have shown, such approximations can be a good place to look.

The reason for this, I want to argue, is that in some cases, *the procedurally, or boundedly, rational solutions to the underlying algorithmic problem are close enough to the substantively rational solution to the computational problem* that such approximations are exactly the right place to look for possible algorithms. In such cases, the rational analysis strategy can be carried directly from the computational level to more algorithmic levels of analysis. This is more difficult than rational analysis at the computational level—both to identify the appropriate kinds of bounds and the space of possible algorithms, and to determine what is optimal or rational given such assumptions—but, to use Anderson's words, I believe it is an even "higher-risk and higher-gain enterprise" than a rational analysis that remains at the computational level. The recent wave of work on rational process models based on sampling has already been essentially following this path. My goal here is to begin to explain the sense in which this work follows from the general rational analysis strategy, thus offering a partial answer to objection (3) from §3.1.

Contrary to more orthodox construals of rational analysis, I believe that boundedly rational analysis can and should make use of what we know about how brains work, and anything else that may be relevant to determining the relevant costs.[3]

[3] This echoes a number of recent discussions (see, e.g., Danks 2008; Danks and Eberhardt 2009; Jones and Love 2011).

As Churchland (1986) commented early on in reaction to Marr (1982), "top-down" strategies in psychology already seem to allow "downward glances" to suggest ideas and constrain computational-level hypotheses. I take the example from Movellan and McClelland (2001), which is thoroughly based on brain-inspired models, to illustrate the fruitfulness of combining top-down with bottom-up strategies. As we will see, comparing the bounded rationality of Anderson's MAP method with the particle filter model may also require knowing something about neural cost. Of course, we can still remain neutral about most aspects of anatomical detail with a boundedly rational analysis. But anything we do know should be used in the analysis, if it can help understand costs and what would turn out to be boundedly rational for agents with such characteristics.

3.5 A Sketch of a Theory of Bounded Rationality

We will model an agent's environment using a prior probability distribution $P(H)$ over latent states of the world H, together with a likelihood function $P(D_1, \ldots, D_n|H)$ for sequences of n observations. Thus, upon making the first n observations $\mathbf{D} = D_1, \ldots, D_n$, once again the posterior probability for H is given by Bayes Rule:

$$P(H|\mathbf{D}) \propto P(\mathbf{D}|H)\, P(H)\,.$$

Our agent will face a decision problem, with some set $\mathcal{A} = \{A_1, \ldots, A_m, \ldots\}$ of possible actions, and a utility $u(A_i, H) \in \mathbb{R}$ for all $A_i \in \mathcal{A}$ and each value of H. Let us call the initial distribution, a sequence of observations, and a decision problem together a *scenario*.

This describes our view as theorists looking in at an agent's situation. The probabilities are based on our own estimates, and the utility function is based on what we think the agent's goals are (or ought to be, which may or may not be the same as what the agent thinks its goals are, if it is sophisticated enough to represent its own goals at all). This covers step 1 and 2 of Anderson's strategy.

Making no assumptions about the agent's computational limitations (thus skipping step 3), we can define an *agent function* α to be a mapping from observations $\mathbf{D}$ to a distribution $\alpha(\mathbf{D})$ over $\mathcal{A}$. That is, $\alpha(\mathbf{D})$ assigns a probability to each $A_i \in \mathcal{A}$:

$$\alpha(\mathbf{D})(A_i) \in [0, 1] \; ;$$

$$\sum_j \alpha(\mathbf{D})(A_j) = 1 \; .$$

The *fitness* ϕ of an agent function α is given by:

$$\phi(\alpha) \;=\; \sum_H P(H) \cdot \sum_{\mathbf{D}} P(\mathbf{D}|H) \cdot \sum_j \alpha(\mathbf{D})(A_j) \cdot u(A_j, H) \; . \tag{3.5}$$

The intuition behind Eq. (3.5) is quite simple: nature chooses a state H and generates some observations $\mathbf{D}$ based on H; then the agent must take an action A_i; the payoff is the weighted sum of utility for each of the actions it might take. When it exists, an optimal (highest fitness) agent function α^* is one that never chooses an action A_i whose expected utility under the posterior distribution (conditioned on $\mathbf{D}$) is dominated by another action A_j:

> Optimal Agent Functions: α^* is optimal if for all A_i and $\mathbf{D}$: $\alpha^*(\mathbf{D})(A_i) = 0$, whenever there is $A_j \in \mathcal{A}$, such that
>
> $$\sum_H P(H \mid \mathbf{D}) u(A_j, H) \;>\; \sum_H P(H \mid \mathbf{D}) u(A_i, H) \; .$$

This notion of optimality (step 4) essentially captures the computational level problem that is assumed to underly many cognitive systems. On Anderson's analysis of categorization, H is the state of the world, which in this case is a specification of all the properties of all the objects in the world. The observations are sequences of objects; then upon viewing a new object, the agent must act appropriately, depending on an unobserved property of this new object. In many cases we can simply assume that the actions correlate one-to-one with the possible values of the unobserved variable (e.g., poisonous : avoid, nutritious : consume), and the utility is positive for a correct guess, and zero or negative for an incorrect guess, for example.

Given that most problems of interest are hard, with the associated optimal agent functions intractable, we want to study not just abstract agent functions, but more concrete representations of agents and the actual computations they perform. Suppose we have fixed some class Π of programs in a given language. We can think of programs $\pi \in \Pi$ as reflecting the mental steps an agent goes through in the course of receiving data $\mathbf{D}$ and deciding which action A_i to perform. Following each new data point D_k, we assume there is some distribution over $\mathcal{A}$ reflecting the agent's proclivities to perform various actions, at that point in time. In this way we can think of π as refining a more abstract agent function α_π. The agent functions partition Π into classes of programs that implement the same function. In this general characterization, we leave open both what the "programming language" is—many computational paradigms are conceivable, including familiar computer programming languages, as well as more biologically inspired formalisms—and how fine the partition on Π is.

We also assume we have some way of assigning a *cost* to computation. This might involve time, space, energy, opportunity cost, or any other factor that seems relevant. One can imagine looking at very precise measures of these cost variables for different types of programming languages. This level of precision may depend on the physical substrate in which the computation is being performed, since varying amounts of energy, time, etc. are used in different physical systems, even in carrying out the same algorithm. Parallel processing may save time; spatially local message-passing can save energy; and so on. To the extent that we know such details about how a given problem is solved in the brain, it makes sense to incorporate them in the determination of costs for programs π (or for that matter, of the set Π of programs). But recall that one of the main uses of rational analysis is in cases where we are uncertain about such details. This is why in step 3 Anderson proposes that one (need only) make minimal assumptions about computational costs and limitations.

Let us suppose, very abstractly, that we can associate with a given program π, under a certain scenario, an expected cost that depends on which, and how much, data come in before decision time: $C_\pi(\mathbf{D})$. The *cost-adjusted fitness* of π is then:

$$\phi(\pi) \;=\; \phi(\alpha_\pi) - C_\pi \;, \tag{3.6}$$

where $C_\pi = \sum_H P(H) \cdot \sum_{\mathbf{D}} P(\mathbf{D} \mid H) \cdot C_\pi(\mathbf{D})$ is the overall expected cost. That is, we simply take the fitness of the program's associated agent function less expected costs.[4] We say an agent is *boundedly rational* to the extent that the cost-adjusted fitness of its program is high.

3.5.1 The Scope of Boundedly Rational Analysis

I propose that a boundedly rational analysis proceed just as a rational analysis would: by making assumptions about the environment (captured by variables H and D_i, and utility function u) and the space of possible agents (as represented by their associated programs $\pi \in \Pi$). Anderson's rational analysis step 4 is to derive the optimal behavior given these assumptions. As remarked, this is usually straightforward for *substantive* rationality or optimal. For bounded/procedural rationality, given any non-trivial class Π, with a sufficiently wide class of possible computations, this can be quite difficult. In a related study on bounded computation, Russell and Subramanian (1995) were able to prove optimality for a very simple mail sorting algorithm. Beyond such simple, circumscribed scenarios, I am not aware of any successful examples. Indeed, it has long been recognized that the search for efficient, effective algorithms in complex domains is very much of an empirical science, gradual and experimental (Newell and Simon, 1976). There is every reason to suspect nature's "search" for solutions is analogous, suggesting that absolute optimality analysis may be of only limited use. Certainly, work in Bayesian psychology has not explicitly presented bounded analyses of Bayesian approximation algorithms, sufficient to show these algorithms are better than all possible alternatives, and that is not what I will be attempting here either. Yet, for sufficiently complex problems that matter to an agent, there

[4] An alternative analysis would be to propose a cost *budget* for a given scenario and restrict our attention only to programs that do not exceed this budget, thereafter ignoring differences in cost (Russell and Subramanian, 1995; Krause and Guestrin, 2009). This would be appropriate, e.g., for agents that are dedicated to only a single task.

are some general characteristics of Bayesian agents which do suggest that algorithms possessing enough of these characteristics will be better off than those that do not, all else being equal. This may allow us to narrow down the possibilities enough so that we can ask more precise questions about bounded rationality, sufficient to find a "local optimum" among a reasonably small set of alternatives.

In the course of the next several sections, I consider potential strategies for narrowing in ever closer on more restricted classes of algorithms, and how these restrictions might be vindicated by considerations of (bounded) rationality:

- Perhaps the first general question is when and why, in the first place, it makes sense for an agent's mental state to be factored into representations corresponding, even roughly, to beliefs and desires or goals.

- On the assumption that an agent can be so factored, one can ask when agents that employ specifically *Bayesian approximation* algorithms will be better adapted than other agents who maintain something like beliefs and desires or goals.

- Finally, if we can justify restricting attention to agents that faithfully carry out approximately Bayesian computations, we still face the question of which among the possible approximations is actualized.

On the question of bounded rationality, we ought to bring to bear any tools that seem useful in characterizing what a rational solution to a given problem would be. In view of the difficulties involved in proving analytical results, we can alternatively drawn on simulation results. In circumscribed cases, simulations give us a very reliable picture of the situation and can be just as helpful as analytical results.

To stress once more, if we can reach the point of establishing even rough correspondences between elements of computations and concrete anatomical structures (as in Yang and Shadlen 2007 on low-level visual decisions), this automatically narrows down our search space. While rational analysis may still play a role in understanding what these processes are for, and in what sense they might be rational, we do not need to invoke the strategy for the purpose of discovering the mind's algorithm. The boundedly rational analysis strategy is intended to be particularly relevant when we have not yet achieved such correspondences.

3.5.2 Calculation versus Look-up

Once we begin considering the relative bounded rationality of quasi-concrete programs π, it can be helpful to separate questions of *representation* from questions of *calculation* or *computation*. A class Π of programs may include quite different kinds of architectures, and thus quite different kinds of representations involved in the agent's computations. Comparing these can be more difficult than holding fixed what representations are being used and simply comparing different ways of computing with those representations. Comparison of Anderson's MAP algorithm with Sanborn et al.'s particle filter algorithm, for example, is of the latter sort. But absent independent evidence that this particular representational assumption is justified—in the categorization case, the assumption that an agent maintains some number of partitions of the observations seen so far, and uses that to make predictions—boundedly rational analysis would ideally allow us to compare these algorithms with algorithms that use totally different representations. Some agents, for example, may score very highly on our measure of bounded rational by minimizing the need for representations.

One of the most basic questions about an agent is whether it maintains representations whose function is to track states of the world, possessing what is sometimes called a *mind-to-world direction of fit*; and whether it also maintains representations with a *world-to-mind direction of fit*, something we might want to call a goal, preference, utility, or other determination of value (recall the discussion in the previous chapter, §2.11). There have been a number of proposals for agent architectures that lack such representations, including: some "fast and frugal heuristics", Millikan's "pushmi-pullyu" representations that blend descriptive and directions aspects (Millikan, 1984), and some more radical proposals that eschew representation altogether (Brooks, 1991). In the reinforcement learning literature, there has also been discussion of two different kinds of learning: model-free and model-based, the latter including a descriptive component, a "model" of the world, and the former merely tracking desirabilities of possible actions (Sutton and Barto, 1998). Both seem to be utilized by the brain (Daw et al., 2005).[5]

[5] In fact, Daw et al. (2005) have argued that the brain rationally trades off between these two, favoring whichever is more likely to suffer less error in a given circumstance. This kind of rational

There is also a sizable literature in philosophy of mind and biology, about why and how minds that track states of the world could have evolved, and what advantage this ability accords (see, e.g., Godfrey-Smith 1998). Our goal here is slightly less ambitious: we would merely like to know when, and in what sense, it could be more boundedly optimal than conceivable alternatives in a synchronic sense. Answering this question is important, because for many types of agents in many types of environments, simpler alternatives like those mentioned above may indeed be better off, if we restrict attention to performance in a single scenario. A particularly compelling example of this is the following, inspired by a discussion in Maloney and Mamassian (2009) on Bayesian analyses of vision.

Consider a point estimation problem in which the underlying state of the world is drawn from a normal distribution $S \sim \mathcal{N}(\mu, \sigma_1^2)$, where μ is the mean and σ_1^2 is the variance. The agent obtains a noisy reading D of S, which is also described by a normal distribution around the true point S, i.e., $D \sim \mathcal{N}(S, \sigma_2^2)$, for some σ_2^2. With action space $\mathcal{A} = \mathbb{R}$, the utility function for making an estimate $\tilde{S}$ when the true value is S is given by the squared error:

$$u(\tilde{S}, S) = -(\tilde{S} - S)^2 \ .$$

The optimal agent function is the one that maximizes fitness according to Eq. (3.5) as given above in §3.5 (minimizing expected error, making the necessary adjustments to Eq. (3.5) for the continuous setting). Once we consider agent *programs*, refining the more general agent function, several possibilities emerge. The agent could separately represent information about the state of the world and about the problem being solved—at least roughly corresponding to descriptive and directive representations—and combine them in some appropriate way. Alternatively, at least in cases like this, it is possible for the agent to manifest the same behavior with a much simpler method. Letting $\tau_1 = {}^1\!/\sigma_1^2$ and $\tau_2 = {}^1\!/\sigma_2^2$—the *precision* of S and D, respectively (DeGroot,

tradeoff between computations is indicative of metareasoning, to which we return in the next chapter.

2004, 38)—the optimal agent function can also be described by the following:

$$\tilde{S} \;=\; \frac{\tau_1}{\tau_1+\tau_2}\mu + \frac{\tau_2}{\tau_1+\tau_2}D \;,$$

as a function only of the data point D. In other words, performing optimally in this task requires merely being able to apply a linear map of the form $x \mapsto a + bx$.

I ran a number of simulations with an agent that learns a and b through simple linear regression, with different settings of the learning parameter, and recorded its performance after varying numbers of learning trials. When the learning parameter is as high as 0.1, it does reasonably well after only 10 trials but soon after levels off in performance, remaining suboptimal. If it is set lower, e.g., near 0.01, it takes much longer to perform well; but eventually, after about 100,000 trials, its performance is indistinguishable from the agent that straightforwardly computes Eq. (3.5) with known mean μ and variances σ_1^2 and σ_2^2. Thus, this "look-up table" agent can learn to behave optimally given enough data. If the scenario in which we are assessing agent fitness is one where training time is cheap and amply available, then this is a case where we would not be justified in assuming that an agent adapted to this setting will implement something approximating Bayesian calculations. There would be no reason for an agent facing this environment to maintain separate beliefs about the world, since it can simply apply a very simple rule to decide what to do.[6]

Maloney and Mamassian (2009) discuss a way of behaviorally testing whether an agent is implementing a simple look-up table method like that above, or something closer to Bayesian calculations with separate representations of prior, likelihood, and value-based information (utility, gain, goal, etc.). The idea is simple. Suppose an agent has learned to perform optimally on two different tasks with different associated probabilities and utilities, but over the same state space: say, P_1 and U_1, and P_2 and U_2. That is, the agent has learned to act as though it is maximizing the P_1-expectation of U_1, and the P_2-expectation of U_2 in the respective situations. If it is

[6] As Lieder et al. (2012) explain, essentially the same normal-normal model applies to the setting of ordinary point estimation problems, where there is little chance that people could have sufficient training experience with novel questions that they may not have considered before, e.g., about the duration of Mars' orbit around the sun. This anticipates the discussion in the next section §3.5.3.

possible to signal to the agent that it is in a new situation involving probabilities P_1 but utilities U_2, and the agent is able to act as though it is maximizing P_1-expected utility for U_2, that would strongly implicate separate representations. An agent that was learning a linear function, such as that described above, would require a new training phase and would thus not show immediate transfer. The same transfer test could be made for uncovering separate representations of prior and likelihood, and Maloney and Mamassian (2009) describe several experimental results that suggest the visual system is sometimes able to exhibit these varieties of transfer.

3.5.3 Transfer, Generalization, and Broader Contexts

Whatever turns out to be the truth concerning the visual system, this idea of transfer applies even more straightforwardly to high-level cognition. Take the problem of categorization, for example. While some cases of property/feature induction may be closely tied to specific decision problems (of the [poisonous : avoid, nutritious : consume] variety), often decision problems vary considerably and present novel challenges. At any moment, an agent may have to make a prediction about some hidden feature for an object possessing observable features never before seen. Methods like the particle filter and the MAP algorithm are well suited for this kind of general-purpose inference, extending effectively to cases where relatively little data has been observed. It is clearly useful to have a model of the world that can be exploited to generate predictions about a wide range of events or objects quickly and efficiently, without having to relearn an entirely new set of associations between observations and actions. People obviously have this ability when it comes to categorization, and the rational basis of this ability should be obvious.

A closely related, but distinct, advantage of maintaining a probabilistic model of the world is that it can dramatically improve learning in novel contexts. One particularly important idea is that of *learning to learn* (confusingly also called *transfer* in much of the literature). In many Bayesian models, learning in one domain can improve and accelerate learning in a different, but related, domain. This idea has proven extremely useful in machine learning and robotics (Baxter, 1997; Thrun, 1998), and has also started receiving attention in cognitive psychology (Kemp et al., 2010), including studies of categorization with hierarchical versions of the Dirichlet process model described above (Canini et al., 2010). It is just one of many advantages related to the "blessing of abstraction" that one is accorded with sufficiently powerful Bayesian models (Tenenbaum et al., 2011).

These ideas, related to transfer, generalization, abstraction, etc., have been successful in modeling aspects of higher level cognition, including areas where no other precise, successful computational models exist. But how should we think of their (boundedly) rational justification? This requires considering not just a single scenario and expected performance in that scenario, but an entire *distribution over sequences of scenarios*. An agent will face many problems, and encounter many different sources of data and information, over the course of its lifetime. It is only at this level of investigation that some of these more general characteristics and advantages of Bayesian methods will be apparent.

In theory, extending the bounded rationality framework sketched above to this more general setting is straightforward. We can use the same class of programs Π, and simply consider their performance over more extended periods, meeting one decision problem after another, with the probabilities of sequences governed according to some distribution assumed to capture the statistics of the environment. This brings to the fore both the fact that an agent meets multiple decision problems, and the fact that there is uncertainty about each individual decision problem. In practice, studying this more general setting is of course difficult, requiring more and more trials to obtain reliable results in simulations. I have not pursued this direction in detail for this dissertation, though such considerations do surface in some of the simulation results discussed below. There is little doubt that these advantages would be even more apparent with further analysis.

To summarize, if we consider only a single scenario, it may look as though look-up table agents, and potentially other representationally or computationally simpler architectures, fare better than agents that maintain some model of the world. Indeed, for agents (or parts of agents) that in fact face such environments, they may well be better off employing such algorithms, and we would not be justified in assuming, e.g., a

Bayesian analysis, is appropriate. For more complex cognitive functions, however, the suggestion is that the advantage of maintaining a model—and a probabilistic model specifically—may lie in the higher-level uncertainty over (sequences of) scenarios. If an agent can use what it has learned about one context to help it in another context, this is obviously an advantage. Importantly, these are advantages that approximation algorithms inherit from the "ideal" models pitched at the computational level.

While the above-cited literature in machine learning and psychology attests to the advantages of generalization, transfer, and other features associated with Bayesian models, we would obviously like more precise results demonstrating these advantages, either analytically or through simulation. We cannot yet claim Bayesian approximations are the only models that could possibly exhibit these important characteristics. This may or may not be the case, and the matter merits further attention. Until alternative models are discovered, Monte Carlo methods, as well as other approximations to Bayesian models, appear among the most promising places to look for the mind's solutions to high-level cognitive problems. For the remainder of this chapter, I shall assume that we have achieved this step in boundedly rational analysis, and therefore restrict attention to single scenarios (for tractability). But I would like to flag this as a significant open problem that needs to be addressed.

3.5.4 Boundedly Rational Categorization

Suppose we have decided—on the basis of (bounded) rationality considerations—that a given cognitive process ought to be modeled by an approximately Bayesian algorithm. This still leaves open a number of possibilities. Here, too, we ought to be able to apply boundedly rational analysis to narrow down the possibilities. To give a concrete example of how this might go, I would like to focus on comparing Anderson's original local MAP algorithm with several alternatives, specifically the particle filter with various numbers of particles. Recall that Sanborn et al. (2010) showed the single-particle-filter algorithm accounts for the data better than the MAP algorithm, which in turn provides a better fir than the particle filter with 100 particles. Could we *retrodict* this result, solely on the basis of bounded rationality?

To estimate the performance of these algorithms in simulated environments, I have run simulations where the environmental statistics reflect our assumptions (or rather, Anderson's assumptions: see Appendix B.1) about the underlying problem being solved. I compared Anderson's local MAP algorithm with four different particle filter algorithms—with 1, 2, 5, and 10 particles—as well as a baseline *reflex agent*, which makes random predictions on new observations, and maximizes with respect to count frequency on previously observed objects, exemplifying a look-up architecture. In all cases, the agent is assumed to see some sequence of data produced by a Dirichlet process, consisting of N objects varying along d binary dimensions, before observing a test object with some feature hidden. A correct guess of the hidden feature gives a payoff of 1, otherwise 0. Representative results are shown in Table 3.1.

	$d = 3$, $N = 3$	$d = 5$, $N = 8$	$d = 3$, $N = 30$
reflex agent	0.561	0.555	0.654
local MAP	0.601	**0.626**	**0.656**
particle filter (1)	0.573	0.603	0.625
particle filter (2)	0.596	0.623	0.629
particle filter (5)	**0.606**	**0.626**	0.642
particle filter (10)	0.608	0.640	0.662

Table 3.1: Average payoffs for categorization agents (coupling parameter $c = 0.5$).

For each setting 10,000 iterations were run. The particle filter with 10 particles always received the highest payoffs. The second highest are highlighted in bold. The coupling parameter c was set to 0.5 (see Appendix).

Making no assumptions about cost, the particle filter with 10 particles has the highest average payoff across all simulations, which is perhaps unsurprising. More interesting is what happens with the remaining algorithms. There are situations in which the particle filter with 5 particles performs better than the local MAP algorithm, situations where the MAP algorithm performs better, and situations where they perform roughly equally well. Thus, optimality of the associated agent functions does depend on the nature of the scenario. In scenarios with long enough sequences of simple enough data, the reflex agent is even among the most successful, in line with the results mentioned earlier for the look-up table agent in the normal-normal model.

It would appear that the single-particle-filter is not particularly well adapted. While it is an approximation to the "ideal" model, in the sense that its asymptotic behavior converges to the ideal calculations, it does not look rational in any sense we (or nature) would care about. In particular, it seems worse than the MAP algorithm,[7] which might be worrisome given that Sanborn et al. (2010) showed it provides a more accurate account of the psychological data. Is this the point where rationality/optimality is no longer a guide, and a rational analysis must end? To explain why not, I shall discuss each alternative in turn.

It is of course no surprise that adding more particles leads to better results. This gives the agent a better chance of narrowing in on the optimal clustering of the data and prevents being misled by unrepresentative data early on. While Sanborn et al. (2010) did not explicitly study numbers of particles between 1 and 100 (A. Sanborn, *p.c.*), it is easy to see that any small number would likely provide an equally good alternative. In many cases, the MAP, $R = 1$, and $R = 100$ algorithms all matched the data well. For the other cases, particle filters with 5 or fewer particles also exhibit order effects, and would predict individual variation. At the same time, if maintaining more particles comes at a cost, this cost would have to be very low for the increase from $R = 5$ to $R = 10$ particles to be worth the small gain in fitness. Simply looking at the marginal increase in (estimated) fitness of keeping two particles instead of one, we see that it would only be worth the extra time, space, and energy if such costs amount to less than 1-2% of a utile (or more generally, 1-2% of the difference between payoff with a correct and with an incorrect prediction). The difference between two and five is yet smaller. Thus, restricting attention to the same particle filter model and merely varying the number of particles, a boundedly rational analysis might tell us that maintaining a single particle is resource optimal, just as taking a single sample can be in one-shot inference problems (Vul et al. 2009, §3.6 below). Using one particle already puts the agent well above chance; perhaps further improvements are not worth the additional costs. This is an empirical question, and a case where

[7] This is as Anderson (1991a) predicted, since the single-particle-filter satisfies his desiderata.

knowing more about neural costs could help settle the matter.

On the face of it, the local MAP agent would seem to confront the same costs as the single-particle-filter agent. Both maintain a single partitioning of the observations at any given time. The only difference is that the MAP algorithm *deterministically* selects the MAP partition, whereas the particle filter *stochastically* samples a partition, which may lead to more mistakes on average. Notably, if the coupling rate is higher, the single-particle-filter can outperform the MAP agent. Running the case of $d = 3$, $N = 3$, and setting the coupling parameter to $c = 0.75$—thereby making old categories more likely—the results in Figure 3.2 show even the single-particle-filter outperforms the MAP algorithm. Again, this is because the MAP algorithm always succumbs to "garden paths" (recall the syntax example ($*$) from §2.8), which occur more frequently with a higher coupling parameter; whereas the particle filter has some chance of avoiding these situations.

	$d = 3$, $N = 3$
reflex agent	0.583
local MAP	0.612
particle filter (1)	0.628
particle filter (2)	0.649
particle filter (5)	**0.655**
particle filter (10)	0.666

Table 3.2: Average payoffs for categorization agents (coupling parameter $c = 0.75$)

But let us suppose, for the sake of argument, that the coupling parameter is lower, and the MAP agent does have a slight advantage over the single-particle-filter. Still, I believe there is good reason to assume the cost of the particle filter is significantly less. The MAP algorithm requires searching through the space of all (extending) partitions after each new data point in order to find the one with maximum probability. One might think of the particle filter as having to carry out the same search and then *adding noise* to make the update stochastic. But this is almost certainly the wrong way to think about it, and points to another place where "downward glances" can help us understand costs. As reviewed in some detail in the previous chapter, there is good empirical evidence that computations in the brain are essentially noisy, and that we should view sampling algorithms such as the particle filter as *harnessing* this noise to the agent's advantage, rather than adding noise to otherwise deterministic computations. From this perspective, it would require further energy and resources to eliminate this noise at each step, to bring us from the particle filter to the MAP algorithm. If this is the right way to think about it, and if these simulations are indicative of typical scenarios, then the cost of reducing noise would have to be less than 2-3% of a utile for it to be worth the effort.

Finally, what about the reflex agent? Computationally speaking, it is quite simple—presumably less costly than any of the other models—yet, with enough data and small enough dimensions, it can perform significantly better than even the five-particle-filter. If one could be certain that the environment and scenario would have this form—a long sequence of low-dimensional data preceding a decision with a small number of alternatives—then I think one could not deny that the reflex agent would be better suited than any of these alternative agents. This is another instance where the nature of the scenario would *not* justify assuming an agent adapted to solve the underlying problem ought to be Bayesian in any meaningful sense. We can note that the reflex agent performs well only in restricted cases, and thus that such an agent would not exhibit the attractive transfer, generalization, or other properties discussed above in §3.5.3. Again, to show this convincingly would require simulations involving uncertainty over sequences of scenarios.

To summarize: under some seemingly reasonable assumptions, the single-particle-filter agent (or perhaps the particle filter with 2-5 particles, depending on cost) is the most boundedly rational from among those considered here. In retrospect, we might have *predicted*, on the basis of bounded rationality considerations alone, that this model would provide the closest fit to human performance. I take this, tentatively, as confirmation of the idea that rational analysis can be extended to *boundedly* rational analysis. Of course, we are still left with the question of where these approximation algorithms come from, and why it would make sense to focus on them in the first place. Assuming progress on this larger question, I offer the results of this section as an example of how we might further restrict the space of options using bounded rational analysis, and in particular using simulation results demonstrating the relative expected utility of alternative agents.

3.6 "One and Done"

We have mentioned several times the results of Vul et al. (2009) and Lieder et al. (2012), exploring resource rationality. In this section we explain how this work fits into the bounded rationality framework proposed here.

Suppose we have an underlying distribution $P(H)$ over hypotheses and data $\mathbf{D}$ with likelihood $P(\mathbf{D}|H)$. We furthermore have an action space $\mathcal{A} = \{A_1, \ldots, A_n\}$ with finitely many actions, and a utility function U. Let the space of possible agent programs Π consist, for each $k \in \mathbb{N}$, of a sampling agent π_k, which draws k fair samples from conditional distribution $P(H|\mathbf{D})$ before choosing an action A_i with the highest summed estimated utility (thus using essentially Decision Rule B from §2.11). If we ignore costs of further samples, focusing on agent functions, then it is clear that $\phi(\alpha_{\pi_j}) \leq \phi(\alpha_{\pi_k})$ whenever $j < k$. But if we assume that every additional sample costs a constant amount C, then we can determine the cost-adjusted fitness of the programs in Π, and even determine the optimal agent, given different assumptions about the relative weight of C and the payoffs from the decision problem, and the probabilities. Vul et al. (2009) have in fact carried out this analysis and showed, perhaps surprisingly, that in many cases the optimal agent takes only a single sample before choosing an action. As discussed in the previous chapter, this has important repercussions for the explanation of probability matching behavior.

Following up on this work, Lieder et al. (2012) explore a setup in which Π consists not of perfect sampling agents, but of agents using a concrete Markov chain Monte Carlo algorithm (Metropolis-Hastings). The question in this context is how much bias should be permitted in drawing any particular sample. They show, again under reasonable assumptions about the cost of eliminating bias by running the Markov chain longer, that resource-optimal agents will tolerate considerable bias. As they discuss, this suggests an explanation of the well-known anchoring effects from the heuristics and biases literature (see §2.9 above for further discussion).

Both of these analyses explain central psychological findings, and importantly, relate these findings to the Bayesian framework, which is crucial for a proper response to (1) and (2). But given our current discussion, they amount to *justifications* or *rationalizations* of the relevant behavior only in as far as we have already convinced ourselves that the only possible programs are based on sampling. Allowing other types of programs to be included in Π may make all of these sampling agents look quite suboptimal. In that sense, they do not tell us why (or when) we might expect sampling-based analyses to be successful in the first place. For that, I claim, we must proceed as outlined in the previous sections.

On the other hand, the realization that different numbers of samples may be more cost-effective for different situations points to an additional advantage of (sampling-based) approximation algorithms: flexibility. We explore this in the next section.

3.7 Flexibility and Metareasoning

In §3.5.3, we focused on characteristics of Bayesian approximations that are inherited from the computational level model. The ability to generalize to novel problems, to transfer aspects of learned structure from one domain to the next, to perform "inductive leaps" given little data, etc., are common to all approximate algorithms for calculating with probabilistic models.

However, there are other advantages of these models, specific to approximations such as those based on sampling, which are only apparent from the point of view of bounded rationality. A particularly attractive feature of sampling-based inference algorithms, for instance, is that they are potentially *anytime methods* (Dean and Boddy, 1988), in the sense that they give a reasonable response at any point in the reasoning process, and only improve their performance when given more time. As Vul et al. (2009) showed, a single sample can often reduce uncertainty enough to improve decision quality significantly. But if more time is available, or if drawing further samples is otherwise inexpensive, estimation can always be improved, in expectation, with more samples. As Vul et al. (2013) also showed (cf. Vul 2010, and the discussion in §2.10), experimental results suggest people are able to adjust the number of samples they draw, strategically depending on the importance of the decision problem relative to the cost of sampling.

Such strategic inferential flexibility is accorded by the implicit representation of a full probabilistic model. Recall from Chapter 2 the discussion of sampling algorithms based on architectures like the Boltzmann Machine. Some of these models are able to encode quite detailed probabilistic information. Taking a single (even biased) sample can be adequate for some purposes, while for others spending more time extracting information may be worthwhile. Be that as it may, this potential flexibility remains abeyant unless the agent has some way of controlling aspects of its processing—such as number of samples drawn—online, at decision time. It would be interesting if the mind had specialized sampling engines for separate purposes, each optimized to draw a different fixed number of samples appropriate for that specific purpose. But it would be more interesting if the mind was able to adjust the number of samples taken for one and the same process, on different occasions, in a nearly optimal way. The results from Vul (2010); Vul et al. (2013), as well as anecdotal evidence, speak in favor of this latter possibility.

3.7.1 Boundedly Rational Metareasoning: Proof-of-Concept

Strategic adjustment of sampling—taken here as a representative example of the flexibility of sampling-based methods—requires that the agent have some cost-effective way of deciding how many samples to draw in a given scenario. The general version of this problem will be explored in more depth in the next chapter. Here I merely want to give a proof-of-concept, showing that it can be boundedly rational, in the sense described in this chapter, to incur some cost in order to estimate how many samples to take. To take a toy example, suppose our agent has a probabilistic model of the world described by the Bayesian network in Fig. 3.1.

Let us further suppose that all our agent can do is draw individual samples from this Bayes net, conditioned on the values of some observed subset of nodes 0-4.

In computer simulations, I randomly (uniformly) drew parameters for the Bayes net, a state and an observation, and then randomly generated a decision problem that

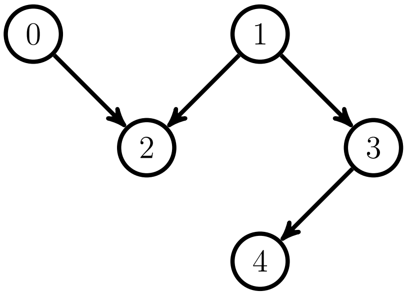

Figure 3.1: Bayesian network for metareasoning simulations.

depended on some subset (possibly all) of the variables 0-4, with utilities ranging between 0 and 100. A *sample cost* C was drawn from a normal distribution with $\sigma^2 = 1.0$ and $\mu \sim \text{Uniform}(0, 3)$. I then compared the performance of eight different agents on the resulting decision problem. The first seven drew fixed numbers k of samples—1, 2, 3, 4, 5, 7, and 9—from the network, conditioned on the observation, and then made a decision using Decision Rule B from §2.11, incurring kC reduction in utility on account of the k samples.

The remaining metareasoning agent first applied a heuristic to determine how many samples to take. This heuristic χ was an extremely simple stepwise function depending only on the cost of a sample:

$$\chi(C) = \begin{cases} 1 & \text{if } 2.5 < C \text{ ;} \\ 2 & \text{if } 1.5 < C < 2.5 \text{ ;} \\ 4 & \text{if } 1.0 < C < 1.5 \text{ ;} \\ 9 & \text{if } 0.5 < C < 1.0 \text{ ;} \\ 15 & \text{if } C < 0.5 \text{ .} \end{cases}$$

The intended interpretation is that χ captures the relative importance of the problem compared to the cost of sampling. Clearly, more sophisticated functions are conceivable, but this is already sufficient to make the point. The results of the simulations are given in Table 3.3.

Importantly, the metareasoning agent suffered the cost of each sample it decided

	average utility
1-sampler	65.6
2-sampler	67.4
3-sampler	67.4
4-sampler	66.8
5-sampler	65.9
7-sampler	63.5
9-sampler	61.1
metareasoner	**68.5**

Table 3.3: Average payoffs for metareasoning agents, out of 100,000 runs.

to take, but it was also charged for the initial step of calculating $\chi(C)$. For the simulation results reported in Table 3.3, the cost of this step was assumed to be equal to the cost of a single sample, which is arguably quite uncharitable for a simple step-function. Nonetheless, the metareasoner still performed significantly better. I also ran 100,000 iterations without charging for the preprocessing step, and in this case the metareasoner's average utility was 70.6. Thus, even if the cost of this simple preprocessing step is constantly 3.0—on average twice the cost of a sample (average sample cost is 1.5)—the metareasoner still outperforms the best constant samplers (the 2-sampler and 3-sampler).

3.7.2 A Distinctive Problem or Implementational Details?

The toy example in the previous section shows how metareasoning, or metalevel control, can improve performance, even when we explicitly consider the cost of that control. This is of course not restricted to controlling the number of samples. There may also be alternative ways of using samples to make a decision, or of drawing samples in the first place. For that matter, sampling is itself only intended as an illustration of a flexible algorithmic method that can be used more or less intelligently; the general point is much broader.

In moving from substantive rationality to procedural rationality, we have characterized the "problem being solved" in terms of a kind of constraint optimization, rather than as a pure inference or decision problem. The problem, we have supposed following Simon (1955, 1956, 1976), is to make a best guess subject to memory, time, and processing constraints. That is, rational use of the agent's own mental resources becomes part of the problem to be solved. Almost dual to the claim from the *extended cognition* literature, that the mind extends out into the environment (Clark, 1997), the claim here is that (parts of) the mind ought to be treated as on a par with the rest of the environment. Indeed, in the next chapter, I will propose viewing the agent's own mental actions or computations on a par with "concrete" actions, so that we can apply the same decision theoretic tools to the agent's deliberative possibilities as to the agent's possibilities for action in the world.

To see how these "internal" decisions are truly analogous to "external decisions", consider the question of how the mind uses samples to make a decision, assuming it does so. We can think of the problem of what to do with samples as a problem of ordinary statistical inference; from this viewpoint Decision Rules A and B end up looking like Reichenbach's (1949) *straight rule*, whereby the estimated probability of a given event is simply its relative frequency. The mind is effectively treating its own generated samples as the statistician would samples of some external process: as indicative evidence about the underlying process. Just as with other cognitive functions, we thus characterize the computational-level "problem being solved" by the mind in regulating the number of samples drawn (or other aspects of cognitive processing) in terms of a statistical inference problem.

All of this raises an important question about the role of metareasoning in the rational analysis strategy. On the face of it, it seems to complicate matters. We have discussed the difficulty of narrowing in on boundedly rational solutions (§3.5.1). The rough suggestion made in §3.5.3 was that bounded rationality considerations may allow us to narrow down the space of possible algorithms by restricting attention to algorithms with good properties; insofar as one can show that any algorithm with all these properties would amount to an approximation to the "ideal" Bayesian model, this would dramatically narrow the search. However, what we have seen in the previous section is that any "pure" sampling agent can be outperformed by an agent that spends time calculating how many samples to take. How could we have known to include this agent in the space of possible agents to consider? Vul et al. did not, for example, even though the results in Vul et al. (2013) suggest that people do seem to be effectively employing some such strategy. This is a relatively simple example, but in as far as the metareasoning agent is more rational, it ought to be considered.

The worry, at this point, is that we have undermined the idea of boundedly rational analysis. Perhaps this example just shows that, once we focus in on messy details of algorithms, the search space becomes too unwieldy, and it is simply too difficult to carry out the rational analysis strategy. Below are three stances one might take toward this issue:

(a) Metareasoning is a matter of implementational/algorithmic-level detail and not properly part of rational analysis. When investigating some cognitive function (memory, language understanding, causal inference, etc.), the possibility of metareasoning will never feature in the computational-level description of the problem, and so can only come in when we are describing possible mechanisms or algorithms. At that point, we are no longer proceeding by rational analysis.

(b) Rational analysis ought to be more fine-grained—more like what we have called boundedly rational analysis—where we are explicitly comparing different solutions to a problem (whether of memory, language understanding, etc.), some of which may explicitly involve metareasoning steps. Thus, instead of comparing solutions to abstract inference problems, we should be comparing concrete algorithmic (possibly metareasoning) agents. The fact that this is difficult does not show that it cannot be a promising method.

(c) Metareasoning is itself a distinctive cognitive function, which we can investigate from a rational analysis perspective. In other words, it deserves a place in the same ranks as memory, parsing, causal inference, and so on, as a separable system, which, together with of all of these other systems, conspires to produce intelligent behavior. In keeping with functional modularity (§3.2), metareasoning is itself a stand-alone module.

My own view is that these are all legitimate. The attitude described in (a) is appropriate when we know absolutely nothing about a cognitive function. This is the first step of Anderson's methodology, and it can help dramatically reduce the search space of possible algorithms, as I have tried to illustrate and explain above. However, once we have made some progress—and perhaps after some "downward glances"—it makes sense to formulate more concrete hypotheses, and metareasoning may well feature in these concrete hypotheses. Then, in accord with (b), we can compare different algorithmic strategies using boundedly rational analysis as we have sketched in this chapter. Of course, the search space at this level is dauntingly large. It is possible we may be able to receive some guidance if we can make some progress on (c). If we can formulate a general *problem of metareasoning*, then we can start to investigate optimal solutions to this problem, which can then feed back into algorithmic-level analyses of arbitrary cognitive systems. In the next chapter, I will explore the idea that the metareasoning problem can be posed and solved (at a computational level) using *value of information*. It may then be appropriate to ask what a boundedly rational solution to the metareasoning problem would look like in concrete cases. If the example above of the stepwise function is any indication, we should expect that metareasoning is an area where very simple heuristic solutions, rather than (even approximately) Bayesian calculations, are more reasonable and prevalent. Indeed, as will become apparent in the next chapter, such a stance is forced upon us on pain of potential regress.

3.8 Conclusion

I have proposed that the rational analysis program, initiated by Marr and Poggio (1976); Marr (1982) and carried further by Anderson (1990), can be profitably extended to a program of *boundedly rational analysis*, proceeding on the assumption that people (or proper parts of people) are *boundedly* or *procedurally rational.* Rational analysis already requires a number of assumptions about the nature of the agent's environment, the kinds of representations employed, and the underlying problem that is to be solved. Boundedly rational analysis adds to this the need for assumptions about the class of possible algorithms we consider, including the relative costs of applying those algorithms. In that sense—that we have more opportunity to make false assumptions—it is higher risk than rational analysis. However, I have argued that it is also higher gain. For reasons familiar from Simon (1955) and many since, people can only be expected to exhibit this more limited, realistic brand of rationality. Keeping that in mind both helps focus our search, and helps to make sense of existing results.

As a concrete example of this claimed payoff, I have outlined a proposal for responding to an increasing number of criticisms of the Bayesian program within rational analysis. Recall the main points of this criticism:

(1) The computations that seem to be required by many Bayesian models are intractable. This makes their application to human cognition questionable.

(2) There is a general problem of how a Bayesian model is confirmed by the data. Typically, subjects show a kind of "posterior matching" instead of what would seem to be the ideally rational strategy of making the MAP guess. On the face of it, this rather disconfirms the hypothesis that subjects are rational.

As we explained in Chapter 2, the hypothesis that people are sampling from probabilistic models shows how Bayesian calculations can be approximated, and in such a way that explains population-level "posterior-matching" behavior. But:

(3) If people are merely approximating Bayesian solutions, then in what sense is their behavior rational at all? Moreover, once we give up on perfect Bayesian rationality, why should we be constrained to search for approximations to the Bayesian solution in the first place? Might not other solutions be better for the task at hand?

(4) How are we to determine when a given problem or cognitive function can be appropriately modeled as (approximate) Bayesian inference? Conversely, when should we conclude that a given phenomenon should not be understood as Bayesian inference?

Indeed, as I have tried to show, there are cases where other solutions are more appropriate to a given task. The *boundedly rational analysis* strategy, as I have sketched it here, assumes that in any particular case, we should zero in on models that provide a good, boundedly rational solution to the whole problem, which includes not just the task at hand, but also the problem of managing and rationing one's computational resources efficiently. To respond fully to (4), we need a more complete account of the space of possible solutions to large-scale, temporally extended problems. I have mentioned some of the attractive features of Bayesian models, viz. generalization, transfer, etc., and some particular to sampling-based and other approximation-based models, viz. flexibility and the possibility of metalevel control. All of this, I have argued, justifies a focus on such methods. But a number of questions remain.

In the next chapter, we will focus our attention on the phenomenon of metareasoning. As explained here in §3.7.2, understanding the problem metareasoning may help to answer some of these questions. Conversely, as will become evidence in Chapter 4, metareasoning is a natural consequence of boundedness, at least for sufficiently sophisticated agents like us, and is a rich topic in its own right.

Chapter 4

Metareasoning and Value of Information

4.1 Introduction

The topic of metareasoning—reasoning about one's own reasoning processes—has come up at many points in this dissertation:

- The idea of mental sampling from implicit, internal probability distributions paves the way for a discussion of *rational sampling* methods, taking only as many samples as seems necessary for the task at hand. As we showed in the previous chapter, online reasoning about how many samples to take can significantly improve performance.

- On one interpretation of the *availability heuristic*, people are using second-order properties of their own internal simulations to guide inference. For instance, the observation that instances of a certain event are difficult to imagine seems to be used as evidence that the event is rare or improbable.

- Given the hypothesized existence of distinct *habitual* (or *look-up table*, or *model-free*) and *goal-oriented* (or *model-based*) learning algorithms, one can then ask the *meta-level* question of how the mind arbitrates between these alternatives. Some work shows our minds trade off in a near-optimal way (Daw et al., 2005).

- The very idea that we are bounded and face the problem of how to use our own bounded resources in a way conducive to meeting our goals suggests the need for metareasoning.

Metareasoning was discussed in the previous chapter in connection with rational analysis and the Bayesian program in psychology. In this chapter, I shall approach the topic in general, using the previous discussion as a useful source of application and illustration.

4.2 Savage's Challenge

The very idea that we could have a formal theory for understanding rational metareasoning has been taken by some as problematic. Standard decision theory is not obviously equipped to tell us anything about the topic. In a 1967 special journal issue on subjective probability, Savage raised what he saw as a formidable challenge to the decision theory that he himself played a seminal role in developing:

> A person required to risk money on a remote digit of π would have to compute that digit in order to comply fully with the theory, though this would really be wasteful if the cost of computation were more than the prize involved. For the postulates of the theory imply that you should behave in accordance with the logical implications of all that you know. Is it possible to improve the theory in this respect, making allowance within it for the cost of thinking, or would that entail paradox, as I am inclined to believe but unable to demonstrate? If the remedy is not in changing the theory but rather in the way in which we attempt to use it, clarification is still to be desired. (Savage, 1967, 308)

In Savage's example, intuitively the person has various options. One could make a guess about that remote digit and bet accordingly. One could take some time to compute the remote digit before making the bet. One might even stop to think which of the first two strategies would be best in this case. Classical decision theory, as formulated by Savage and others, simply says that one should put all of one's money on the correct digit, a recommendation that seems unhelpful to the person who happens not to know what digit that is. As the example shows, if decision theory is to give an account of what one ought to do, holding fixed beliefs and preferences, it should take into account the costs and benefits of figuring out what one ought to do. Since figuring out what one ought to do is just another thing one does, with different ways of figuring accompanied by different costs and potential benefits, the question arises of how best to decide how best to decide what to do. As one goes down this path, there does loom a worry of paradox, or at least of vicious regress. Indeed, a decade earlier in *The Foundations of Statistics* Savage expressed just this worry:

> It might be stimulating, and it is certainly more realistic, to think of consideration or calculation as itself an act on which the person must decide. Though I have not explored the latter possibility carefully, I suspect that any attempt to do so formally leads to fruitless and endless regression. (Savage, 1954, 30)

I see two principal challenges represented in Savage's comments. These two challenges are related, but I believe it is important to distinguish them. The first challenge is essentially the problem of *logical omniscience*: In the scenario Savage describes, how can we model the mental state of the agent who does not know the remote digit of π, yet knows enough basic facts about circles, ratios, etc., so that the identity of the digit is a logical consequence of that knowledge? And if the recommendation from decision theory is that the agent ought to maximize expected utility, what is the distribution over which this expectation is to be taken? This is an important problem, and I will return to it in Appendix C.2.

To my mind, the more pressing challenge is one of how to assess alternative deliberative strategies for real agents. There is no doubt deliberation can be worthwhile. When a decision matters greatly and there is ample time to think about it, careful calculation and weighing of considerations can help achieve better consequences. On the other hand, excessive, meandering, or pointless deliberation is typically to be avoided. When time is scarce or thinking too costly, it may be wise to avoid deliberating at all. So much is platitude. The important point is that for any sufficiently complex agent, there may be a genuine question about how and when to spend time and energy deliberating, including deliberating about deliberating. We would like decision theory to shed light on this problem of when deliberation is likely to be worthwhile. I take this to be Savage's second, and main, challenge: to improve the theory, or at least to change the way we use it, by taking the cost of thinking into account. How can we do this in a way that does not lead to endless regress or paradox? Is decision theory somehow fundamentally incapable of shedding light on the processes underlying rational deliberation?

My first aim in this chapter is to argue that Savage's second challenge can be met, without ever broaching the topic of omniscience. Our discussion of bounded rationality in the previous chapter already paves the way for the analysis in this chapter. By studying agents from an algorithmic point of view, and explicitly modeling costs of computation, we can understand why some deliberative paths might be better than others. As I shall explain, the natural way of couching this analysis is in terms of the *value of information*, which concerns a special case of sequential decision problems. However, unlike with most uses of value of information, which are about external observation or experimentation, the theory must be adapted to the setting where information derives from internal computation or deliberation. This analysis works both at a very low level in explanations of automatic psychological processes, and at a high level in explicit, temporally extended deliberative episodes, as I will illustrate with several examples.

My second aim is to argue, as Savage suggested, using decision theory to model bounded agents does require some care in how the theory is interpreted. In short, my response to Savage's worry is that we cannot treat decision theory unambiguously as providing normative guidance about what to do in the context of deliberation. Assuming otherwise immediately leads to contradiction, as the following are obviously inconsistent: (1) In any choice situation we can consult our theory to decide what to do; (2) Consulting the theory amounts to deliberation; (3) There are situations in which the theory recommends not to deliberate at all. I take it (2) is obvious, and (3) should be a simple consequence of any decision theory that weighs deliberative strategies. So we must give up (1).[1] Importantly, this does not mean we have to give up using decision theory in deliberation. I hope this will be clear from the first half of the chapter. But it does mean that we must be explicit about the use of the theory in any particular case, especially if we want to derive any normative implications. Thus, in the second half of the chapter I will discuss three uses of the information value account and how they relate: in designing agents, in rational assessment, and in the context of deliberation. In Appendix C.2 , I will briefly discuss the problem of omniscience and the sense in which it is avoided.

4.2.1 Hacking's Reply

In the same journal issue as Savage's paper, Hacking (1967) offers a reply to Savage's challenge. The paper introduces a number of important issues,[2] but I want to focus on a particular aspect of his reply, as it forges a noteworthy link with our discussion in the previous two chapters.

Hacking discusses the fact that probabilistic models often used in physics, biology, economics, psychology and other disciplines are sufficiently complicated that analytic methods for working with them are either unknown or take a long time even for high-speed computers to carry out. Consequently, scientists will often use Monte Carlo sampling methods and related techniques to estimate the quantities in question. In fact, the original use of Monte Carlo techniques was to estimate integrals that were otherwise difficult to compute. (See Chapter 2, and for an historical account of Monte Carlo methods see Eckhardt 1987.)

Hacking raises this as a perfectly realistic and practical example of where an

[1] Though I do suspect something like assumption (1) underlies the worry about paradox or regress, it is also clear that Savage took decision theory to be first and foremost a theory of consistency and coherence. As he explains in Savage (1954), "When certain maxims are presented for your consideration, you must ask yourself whether you try to behave in accordance with them, or, to put it differently, how you would react if you noticed yourself violating them" (p. 7).

[2] It is actually most famous for reasons tangential to metareasoning. The paper is often cited as containing the first suggestion to use impossible worlds in modeling non-omniscient agents (see §C.2 below). It is also cited for its challenge to Dutch book arguments as a justification for coherence (though see, e.g., Skyrms 1990, Ch. 5).

excessively punctilious application of decision theory would give the wrong result.[3] It is obvious that the data scientist who uses Monte Carlo methods to estimate some value of interest is making the rational choice. If computing an exact value would take a week (or worse, would be intractable) but computing a reasonable estimate would only take several hours, it may well be worth the loss of precision to obtain an approximation and move on to the next task. That is so even if one would prefer an exact prediction to an approximate prediction, all else being equal. The point is that all else is not equal, and the gain in accuracy is not worth the cost of waiting a week more. Furthermore, it is often worth taking the extra time to figure this out, since if an exact calculation can be made, that would be preferable. It may be that a simple calculation, or observation about the form of the distribution, would be sufficient to determine whether exact calculation would be feasible. Hacking's example thus also involves a very simple case of deciding how to decide what to do.

This all seems perfectly obvious, and Savage would surely agree that the scientist is doing the right thing. The challenge is to capture this in the formalism of decision theory, and show that this behavior is indeed rational according to the theory. Hacking made some proposals, but in the end his suggestion amounts to giving up on formalization. As a reply to the worry of paradox, he says:

> I do not find the regress paradoxical. You can allow for the cost of as much thinking as you like, up a long string of meta-metas. But you have to disregard the cost of thinking out the ultimate meta-decision. [. . .] For the computer programmer who arranges a Monte Carlo solution rather than an exact one [. . .] his meta-thinking may take ten minutes of pencil time, while the object level thinking may take hours of computer time. It makes good sense to forget about the pencil time. All practical Bayesian business has to round off estimates of costs to one or two per cent. The cost of meta-thinking gets rounded off. (Hacking, 1967, 324)

Again, this sounds right as far as it goes. But how do we know the cost of meta-thinking "gets rounded off" appropriately? In the case of the scientist using Monte

[3] Gaifman (2004) also cites Monte Carlo integration as an example of reasonable use of probabilistic methods for mathematical investigation. I consider Gaifman's response in Appendix C.2.

Carlo approximation, the cost can be rounded off simply because there is only one level of metareasoning and it is of relatively low cost. How are we to know that, in general, by going up the meta-level hierarchy one can disregard the further costs? In other examples this may well be false. We would like to have a theory that sheds light on when it does hold.

4.3 Value of Information

The necessary theoretical tools to handle these kinds of scenarios were in place long before this exchange, though the crucial connection to cost of thinking and computation had not yet been made. In the early 1960s statisticians and control theorists were interested in understanding when gathering more information, for instance by performing experiments, would be worth the cost of performing the experiments plus the possible costs of delaying action. The classic sources are Raiffa and Schlaifer (1961) and Howard (1966), though the idea was developed independently in the Soviet Union by Stratonovich (1965). As Skyrms (1990) points out, many of the crucial technical observations were already made by Ramsey in the 1920s, and later by Lindley (1956), Good (1967), and by Savage himself (Savage, 1954, Ch. 7). The basic idea of *value of information* is quite simple and can be illustrated by example.

Suppose I am at the train station on my way to an appointment. Two trains are on the track and I am unsure which is going to my destination. If I pick the right train I will make it to my appointment on time. If I take the wrong train I will miss the appointment altogether. A schedule is printed upstairs, but by the time I make it there and back both trains will have left and I will have to wait until the next train in half an hour. I will then be late for the appointment, though at least I will not miss it altogether. What should I do?

Say my utilities for the possible outcomes are U_1, U_2, and U_3, for arriving on time, arriving late, and missing altogether, respectively; and the probability I assign to train 1 being correct is p. Then the expected utility of checking upstairs is simply U_2, since I will certainly make it, but will certainly be late. The expected utility of guessing now is $pU_1+(1-p)U_3$, since there is probability p I will get it right and arrive on time, and $1-p$ I will take the wrong train and miss it. The value of the extra information, ignoring the cost of obtaining it, is then the difference between these expected outcomes: $U_2 - (pU_1 + (1-p)U_3)$. It is worth obtaining the information just in case this value is greater than 0, i.e., when $U_2 > (pU_1 + (1-p)U_3)$.

It is a theorem of decision theory that a more informed choice is always at least as good as a less informed choice (Good, 1967; Skyrms, 1990).[4] That means the value of obtaining further information can always be weighed directly against the cost of obtaining the information, as in this simple example. In the remainder of this section, we develop the general theory of value of information, drawing on previous treatments (Bernardo and Smith, 1994; Skyrms, 1990; Hay and Russell, 2011).

4.3.1 Informational Decision Problems

The general scenario to be analyzed involves a probability model with:

- Random variables $U_1, \ldots, U_n$, taking on values in $\mathbb{R}$, with U_i representing the utility of concrete action i. So, for example, $P(U_i = x)$ is the probability that action i has utility x.
- Random variables $E_1, E_2, \ldots \in \mathcal{E}$, where E_j represents the outcome of performing experiment j. Thus, experiments and concrete actions are kept distinct.

For notation:

- A sequence $\mathbf{E} \in \mathcal{E}^*$ of experiments gives rise to an observation sequence $\mathbf{o}$. $\mathcal{O}(\mathbf{E})$ is the set of all such sequences. The set of all observation sequences for all experiment sequences is given by $\hat{\mathcal{O}} = \{\mathbf{o} \in \mathcal{O}(\mathbf{E}) : \mathbf{E} \in \mathcal{E}^*\}$.
- Let us write $P(\mathbf{o})$ instead of $P(\mathbf{E} = \mathbf{o})$, where $\mathbf{E}$ is implicit.

[4] It is important that the result of obtaining the information does not change the choice situation. For instance, in games one can be at a disadvantage with more information, e.g., if it is commonly known that this information has been obtained. Having certain kinds of knowledge may even give negative utility in and of itself. Kadane et al. (2008) explore a number of interesting cases where the theorem does not hold, including situations in which an agent has an improper or not countably additive prior, as well as situations in which an agent's beliefs are represented by sets of distributions. In many of these situations a rational agent would actually pay not to receive information.

Suppose we perform a finite sequence $\mathbf{E}$ of experiments. Then the expected utility now of acting optimally after having observed the results will be:[5]

$$u(\mathbf{E}) = \sum_{\mathbf{o} \in \mathcal{O}(\mathbf{E})} P(\mathbf{o}) \max_i \mathbb{E}(U_i|\mathbf{o}). \tag{4.1}$$

The utility of acting without performing any experiments is then a special case:

$$u(\epsilon) = \max_i \mathbb{E}(U_i).$$

In other words, without any additional information the best we can do is maximize expected utility. By performing experiments we can obtain new information to improve the utility estimates. For instance, in the train example I may be unsure about what information I will find on the schedule, yet even before I receive the information, I can determine that either outcome will improve my ability to make the right choice. This means the current expected utility of acting after receiving this information may be greater than the current expected utility of acting now.

The probabilities on outcomes should ideally satisfy a kind of reflection principle:

(R) For all U_i and $\mathbf{E} \in \mathcal{E}^*$:

$$P(U_i) = \sum_{\mathbf{o} \in \mathcal{O}(\mathbf{E})} P(\mathbf{o}) P(U_i|\mathbf{o}).$$

That is, the agent's beliefs about how the utility of an action will change after performing experiments should be properly incorporated into the current estimate of utility. This of course does not preclude believing that learning more will change the expectation. But one cannot know beforehand how it will change. This is a special case of Condition (M) from (Skyrms, 1990, Ch. 4), where it is claimed that an agent whose expectations violate this principle thereby expects the action to be "brainwashing, delusional, or otherwise epistemologically pathological" (97). Assuming (R) holds, it is not hard to show that in general $u(\mathbf{E}) \leq u(\mathbf{E}')$ whenever $\mathbf{E}'$ extends $\mathbf{E}$.

[5] This is assuming E_i can take on finitely many values. Otherwise the sum would be an integral.

This is a restatement of the fact that a more informed decision is always at least as good as a less informed decision.

The final ingredient of the model is the cost function $c : \mathcal{E} \to \mathbb{R}^+$. Assuming that $c(E_j)$ is a constant for each j, and $c(\mathbf{E})$ is the sum of those costs, the net expected value of performing experiments $\mathbf{E}$ will be:[6]

$$V(\mathbf{E}) = u(\mathbf{E}) - c(\mathbf{E}). \tag{4.2}$$

Let $\mathcal{M}$ be the probability model together with the cost function.

An agent need not choose a full sequence of experiments right at the beginning. Which experiments to perform in the future, or whether to perform any further experiments at all, may depend on the outcomes of previous experiments. A *strategy* is a function $\sigma : \hat{\mathcal{O}} \to \mathcal{E} \cup \{\mathsf{stop}\}$, assigning an experiment or stop to every possible observational state. Once stop is chosen, the agent performs no further experiments and selects the action with highest expected utility in the state at that time.[7] A strategy σ therefore induces a distribution $p_{\mathbf{E}}$ over sequences $\mathbf{E}$ of experiments (see Appendix C.1 for details). Assuming σ terminates with probability 1, the expected value of σ is:

$$V(\sigma) = \sum_{\mathbf{E} \in \mathcal{E}^*} [p_{\mathbf{E}} \cdot V(\mathbf{E})].$$

This is just the summed probability of performing the sequence $\mathbf{E}$ times the net expected value of performing $\mathbf{E}$.

The *informational decision problem* for $\mathcal{M}$ is to find σ for which $V(\sigma)$ is relatively high, or ideally, maximal. Extensions of this setup, e.g., where we allow sequences of concrete actions punctuated by or even in parallel with experimental actions, are certainly of interest. For our purposes, the special case of a single concrete action proceeded by some number of experiments is sufficient to illustrate the central points.

The informational decision problem is a special case of a sequential decision problem. In fact, this kind of problem is equivalent to that of finding an optimal policy in a corresponding Markov Decision Process (MDP). A verification of this fact can be

[6] Cf. Eq. (3.6) from §3.5

[7] Other assumptions are possible here, e.g., a softmax rule (see §2.10 or Sutton and Barto 1998).

found in Appendix C.1. Recall an MDP is given by a set of states and a set of (possibly costly) actions, where the distribution over next states is completely determined by the action and the previous state (the Markov property). In short, the states in these MDPs are *information states*, identified by vectors of probabilities over the *concrete action* utilities, and the actions are *computational actions* which can shift these probabilities.[8] This equivalence will allow us to visualize an informational decision problem in a simple and perspicuous way. Hacking's Monte Carlo scenario provides an easy example.

4.3.2 Hacking's Monte Carlo Example

An MDP for performing experiments defines a space of possible sequences of experimentation before performing an action. It is most obviously applied to concrete data collection methods, or making concrete observations as in the train schedule example. But the basic mathematical apparatus can be applied to *simulating* observations just as well.[9] Hacking's example of the simulating scientist can be modeled by an MDP, in a fairly obvious way at that.

To fill in some more details, suppose our data scientist needs to estimate the probability of ice on the road, to report to local authorities so that they may take preventative measures by using anti-icing chemicals. If the probability is below a certain threshold, it is not worth using the harsh chemicals which harm the environment; above the threshold, it may be considered worthwhile in order to prevent either a rise in automobile accidents or significant economic losses. The probabilistic model for ice on the roads can sometimes be used to calculate exact probabilities within a few hours, but an hour of calculation is required to determine whether this will be possible. The sooner the authorities know what to do, the sooner they can begin with preparations. We can picture the scenario very schematically as in Fig. 4.1 below. Boxes are choice points and contain the results of investigation so far.

[8] It is important not to confuse the *concrete actions* which end the process and bare some true utility, and *computational actions* which are used to help estimate the utilities of the concrete actions.

[9] A recent trend in philosophy of science is the idea that simulations are a distinctive kind of experiment. Humphreys (1994) discusses the use of Monte Carlo methods in statistical physics.

Outgoing arrows are labeled by (negative) costs. Circles are labeled by the experiment or computation just performed, with outgoing arrows labeled by probabilities of observations.

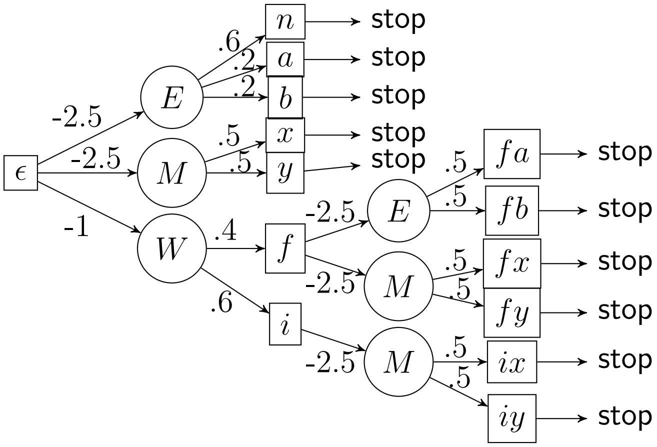

Figure 4.1: MDP for Hacking's simulating data scientist

E is the action of attempting exact calculation, M is running a Monte Carlo simulation, and W is waiting to work out whether exact calculation is feasible. Meanwhile, n is a result showing that exact calculation does not halt, a is the result that exact calculation halts with an estimate above the threshold, b is the exact calculation giving an estimate below the threshold, x is the approximate simulation giving a result above the threshold, y is the simulation giving a result below the threshold, f is a proof that performing E is feasible, and i is a proof that E is infeasible.

For simplicity, suppose the only possible actions are reporting that it will be above the threshold (1) and reporting below the threshold (2). Let us say reporting that it will be icy when correct is assigned utility 15, and when it is incorrect -20. Reporting that it is not icy when it is leads to utility -25, and when it is not 10. Let $P(U_1 = -20) = P(U_1 = 15) = 0.5$; and $P(U_2 = -25) = P(U_2 = 10) = 0.5$. With no further information, the rational thing to do is to take precautionary means and report that it will be icy. This has less disastrous consequences if it turns out to be incorrect, and more positive consequences if it is correct.

$$V(\epsilon) \;=\; \max_{i=1,2} \mathbb{E}(U_i) \;=\; \mathbb{E}(U_1) \;=\; -2.5$$

Suppose the relevant conditional probabilities given possible observations are:

$$P(U_1 = 15 \mid E = a) = .9 \qquad P(U_1 = 15 \mid M = x) = .75$$

$$P(U_2 = 10 \mid E = b) = .9 \qquad P(U_2 = 10 \mid M = y) = .75$$

$$P(E = n \mid W = f) = 0; \qquad P(E = n \mid W = i) = 1;$$

It is easy to check that this model satisfies principle (R). We can now calculate the values of other possible paths. For example,

$$V(E) = u(E) - c(E) = (-1.5 + 2.3 + 1.3) - 2.5 = -.4.$$

$$V(M) = u(M) - c(M) = (.75 + .625) - 2.5 = -1.125.$$

Between risking the exact measurement and taking an estimate, the scientist would rather attempt an exact calculation. Either is better than just guessing. However, if we calculate the value of the strategy σ that first takes W, calculating whether E is feasible, and then takes E if it is, M otherwise:

$$V(\sigma) = (.6 \cdot -1.125 + .4 \cdot 6.5 - 1) = .825$$

This value is greater than either E or M, and in fact it is the optimal strategy for this MDP. So in this situation the scientist should calculate which of the two methods to use before proceeding. In Hacking's example, the calculation shows that exact computation is infeasible, so the scientist will end up choosing W and then M. Assuming this is a reasonable model of the scientist's decision situation, we have shown that the scenario can receive a perfectly standard, and indeed quite simple, analysis in terms of expected utility maximization in a sequential decision problem.

Notice that it would be easy to alter the model, e.g., by dramatically increasing the costs of computation, so that the rational thing to do is simply to act based on the prior, that is, just to report that the roads will be icy. This would be appropriate if, for instance, the authorities needed to take action immediately, with no time for calculation.

4.4 Dynamic Probabilities and Bounded Agents

The point of the above analysis of the informational decision problem is to understand structural aspects of how one should optimally spend one's time gathering information before making a decision. Taking actions in these MDPs changes the information state of the agent, subject to a cost, though this cost may be worth the gain in expected utility afforded by the new information.

In the simulating scientist example the information-producing actions include carrying out proofs and running computational simulations. It is a small step from these still external computations to incorporating in the model computations or procedures internal to the agent. To be sure, one's beliefs and dispositions can change merely through thinking. This is essentially the crucial insight of I. J. Good: the same value of information analysis applies when the information gathering actions are purely internal steps of computation or thinking (see Good 1983 for many of the relevant papers, and Skyrms 1990 for discussion). Indeed, insofar as Savage's gambler has a prior belief about a remote digit of π, which can change in light of further thinking and computation, we could draw a simple MDP to model that situation as well, showing that it may well be rational to take a mixed bet rather than try to compute the remote digit. Subjective probabilities which can change through thinking and reflection are sometimes called *dynamic probabilities*, following Good's terminology.

In the informational decision problem outlined above we may think of the prior probabilities over action utilities as unreflective guesses about how good the outcome of an action is likely to be. The posterior probabilities capture an agent's view of the situation "in the light of" implicitly known facts made explicit, or predictions made on the basis of known facts. These two utility estimates may differ merely because some relevant fact needed to be extracted from memory in order to see what the best action is. They may also differ for a more interesting reason. We often have to make decisions in and about entirely novel situations, different from previously experienced situations, though perhaps similar in some respects, a point stressed in Chapter 3. One could even argue this is what deliberation is primarily for, since when things are familiar or routine we can rely on habit, or imitating past actions that led to desirable outcomes. In any case, it should be uncontroversial that our assessments of utilities—interpreted as dispositions to act—can change merely through thinking. The price of this ability is that deciding the best course of action can take time. Often one must *do* something to make a good decision. A simple example may help to motivate this.

Suppose I need to decide which of two highways to take to some destination. If forced to choose immediately, my choice will be more or less random. However, given some time to think about it I can bring various considerations to bear on the decision: I can estimate how crowded each is likely to be, e.g., on the basis of past experiences, including information about the day of the week and the time of day; I can attempt to recall whether there is construction on one or the other; I can think about whether I find driving on one of them simply more enjoyable; and so on. The more I think about such things, generally speaking, the more likely I am to make a good decision (or at any rate, what I would reflectively consider a good decision given my information). Furthermore, there are better and worse ways of spending this time thinking. Perhaps thinking about likely construction sites is relevant, but less important than traffic. There is an obvious sense in which I do not know how my estimate of traffic will turn out until I actually think about it and make the estimate. Otherwise I would not have to make it. But I can know beforehand that thinking about traffic is likely to be rewarding in this situation. It should thus be clear that the informational decision problem is just as applicable to such deliberative strategies as to experimental strategies.

4.4.1 Questions to Nature and to Oneself

Galileo famously suggested that experimentation can be thought of as posing *questions to nature*. More generally, the actions in the original formulation of the informational decision problem can be thought of as questions of one sort or another. Hacking's simulating scientist poses a question to the computer, for example.[10] If the elements of $\mathcal{E}$—originally conceived as experiments—are questions, then the possible observations $\mathcal{O}(E_j)$, for $E_j \in \mathcal{E}$, can be seen as the possible *answers* to these questions. Different answers will obviously lead to different assessments of the possible actions, making the random variables U_i dependent on variables E_j, in general. Indeed, the theory of value of information has been applied directly to the analysis of questions in natural language (van Rooij, 2004).

Once we interpret the actions as purely internal, thinking or deliberating activities, it becomes natural to think of $\mathcal{E}$ as consisting of possible *questions to oneself*, and $\mathcal{O}(E_j)$ as the possible resolutions to such questions that our minds return. From the scenario described above, in thinking about traffic I can be described as asking myself a question about how the traffic will be. Before I pose this "internal" question, there is a sense in which I do not know the answer (cf. §1.2). That sense is captured in the model by its not being reflected in my dispositions for action, that is, in the random variables U_i. Thus, even for facts implicitly known, e.g., based in memory, it may be necessary to *do* something in order to use that knowledge. I think—or pose questions to myself—precisely so that I can act "in light of" answers to those questions that I have the ability to provide.[11]

This idea of acting "in light of" implicitly known information is the crucial new idea beyond the *bounded rationality* framework outlined in the previous chapter. We now assume that an agent program (in the sense of Chapter 3) involves estimating the desirability of possible actions—this could be achieved by any of the sampling-based decision rules discussed in Chapter 2, for example, but also by many other types of

[10] See also Newell and Simon (1976): "Each new machine that is built is an experiment. Actually constructing the machine poses a question to nature; and we listen for the answer by observing the machine in operation and analyzing it by all analytical and measurement means available" (114).

[11] It is noteworthy in this connection that experimental results show students often perform better in problem solving when they *literally* ask themselves questions in the process (Chi et al., 1989).

programs—and deliberative strategies are compared with respect to how well they sharpen these estimates, by the agent's own lights. The value of a question, or series of questions, is measured by the value of information it is likely to provide about what the agent ought to do, again by the agent's own lights.

Just as with experimentation and scientific investigation, different lines of inquiry—or sequences of questions to oneself—may be better than others. As a special case of the costs associated with programs in our discussion of bounded rationality, there is always a cost to pursuing one path rather than another, if only the opportunity cost lost by not pursuing the other path. That makes salient the meta-level question of which path to pursue, which itself can be posed in different ways, leading us further up the metareasoning hierarchy. There is always a potential higher-level question about what lower-level question one ought to pose. We will discuss the apparent threat of regress below in §4.4.5. First, I would like to consider some examples from cognitive psychology and from artificial intelligence.

4.4.2 Sampling as Thinking

The Sampling Hypothesis provides a perfect setting for illustrating rational metareasoning. As I have argued, sampling makes intuitive sense as a representation of the process of thought. Consider the example above of the decision between two highways. My model of the relevant aspects of the world is clearly quite complex. Anything from weather to sporting events could bear on my utility estimate. In line with Chapter 2, we might think of a sample as a single traversal of this model, picking up relevant considerations and updating the utility estimates accordingly; and in line with the previous discussion, we can think of drawing a sample as posing a question to one's own mind, with answers generated noisily. While we may of course be interested in smart ways of *directing* these samples/questions so that they are maximally helpful, let us suppose for now that the samples come from some single underlying probability distribution on states of the world. States are then weighed as being more likely the more they are sampled, in accordance with the decision rules from Chapter 2. Again, the more samples taken, the better the estimate of (my implicit model of) the true state, and thus the better the action is likely to be.

With sampling now standing in for thinking, the question becomes: how many samples from the distribution should be taken before acting? Since samples presumably come with a cost, this makes the problem perfectly suited to formalization as an informational decision problem, and it also nicely illustrates the use of Good's dynamic probabilities in a precise setting.

Suppose we have two actions a and b, let U stand for the statement that action a has higher utility, and let p be the true probability that $U = \mathsf{T}$. Perhaps p is derived from some complicated model and cannot be easily computed, such as in the highway example. At any point, the agent can either sample p, which in the two-alternative case we can think of as flipping a coin with weight p, or choose stop, at which point action a or b is taken according to whether U has estimated expectation greater than 0.5. The agent always begins indifferently with $P(U) = 0.5$, so before taking any samples an action is chosen at random. After taking N samples with outcomes $\mathbf{o}$ (a sequence of values T and F), the agent estimates p simply by taking a finite approximation:

$$P(U|\mathbf{o}) = \frac{\text{number of times } \mathsf{T} \text{ comes up}}{N} .$$

By the Monte Carlo Principle $P(U|\mathbf{o})$ will converge to p as N goes to infinity (cf. §2.7). For any finite number of samples, the probability of choosing action a is:

$$q = 1 - \Theta(\frac{N}{2}, p, N) ,$$

where Θ is the binomial cumulative density function and $\frac{N}{2}$ is rounded down to the nearest positive integer.[12] If the cost of a sample is c, then we can view the problem of how many times to sample as equivalent to finding an optimal policy for an MDP, as in Fig. 4.2.

Because it is assumed that the agent can sample p perfectly, the space of strategies collapses into the space of unconditional strategies. A strategy is specified simply by

[12] Recall $\Theta(k, p, n)$ is the probability of getting at most k heads in n tosses of a coin with bias p.

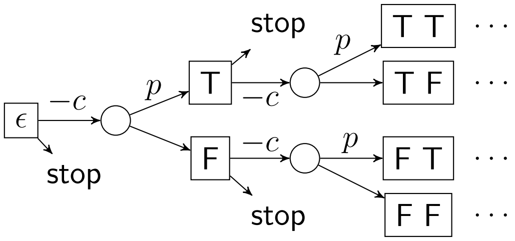

Figure 4.2: MDP for a sampling agent

the number of samples to be taken, where this is unaffected by the outcome of previous samples. In this setting we can interpret the results from Vul et al. (2009) (recall §3.6) as showing that, under reasonable assumptions about the utilities and the cost, the optimal strategy is often to take a single sample. The generalization to arbitrarily many actions is obvious.

It is also worth pointing out that this gives us another demonstration of cases where the thing to do, according to our decision theory, is to act immediately, and in particular not to think at all about what to do. Formally, all we need to do is make the costs in Fig. 4.2 sufficiently high, so that the agent should not draw any samples and take an action immediately. There are also intuitive examples of this. If I am heading at top speed right to a fork in the road with the same choice between highways, it would be unwise to weigh considerations about which highway would be better, or to imagine different scenarios in my head. I ought to act immediately. Again, the framework easily captures this.

4.4.3 Reasoning about Sampling

In describing Fig. 4.2, we the theorists are reasoning about a sampling agent and determining how many samples is optimal for that agent. But recall from the previous chapter that, in order to take advantage of the potential flexibility of sampling, the agent must be able to determine in any given circumstance roughly how many samples ought to be drawn. We showed that an agent who uses a simple stepwise function χ, depending only on an estimate of the relative importance of the problem and the cost of sampling, can outperform any agent that draws a fixed number of samples. We can depict this extended situation—effectively expanding the set of possible computations $\mathcal{E}$ to include computing χ—as in Fig. 4.3. Again, the generalization to multiple possible actions is straightforward.

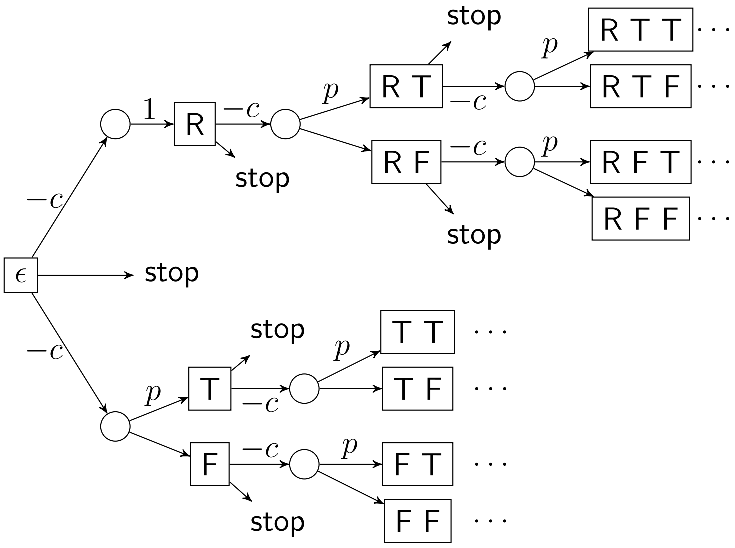

Figure 4.3: MDP for metareasoning about sampling

The downward transition from the initial empty information state ϵ represents taking a first sample without any metareasoning. The upward transition represents computation of χ, which (deterministically) outputs some number R. From that point the MDP looks exactly like the original MDP from Fig. 4.2. An agent could adopt a strategy that ignores the output R of the first computation, or perhaps could use it in an arbitrary way. A smarter agent will proceed as though it were a fixed R-sampler. The results from the previous chapter can be interpreted as providing strong evidence that the optimal strategy in this MDP—for many specifications of probabilities, costs, and utilities, and especially when there is uncertainty over utilities—is to compute χ and use the output of that computation to decide when to stop thereafter.

In this case, the agent need not be explicitly computing the value of information. That is, the strategy that takes the upward transition to compute χ is not directly determining the optimal strategy in the associated MDP depicted in Fig. 4.2. However, χ does approximate that calculation. In that sense, computing χ is a meta-level action: it is a computation *about* possible computations.

If people do strategically adjust how long they think about a given problem—adjusting how many times to sample an internal model, as suggested by Vul et al. (2013)—then they must be effectively solving something very close to the MDP from Fig. 4.2. What analysis of the MDP in Fig. 4.3 shows is that such a metareasoning strategy is indeed justified in terms of the underlying informational decision problem. In the space of possible agents that includes the metareasoning agent and all of the constant-number samplers, the metareasoner is optimal.

4.4.4 Metareasoning and Search

Ideas similar to our metareasoning example for number of samples have been pursued in computer science. For instance, Lagoudakis et al. (2001) show that a metareasoning heuristic that takes as input the size of a list, and deterministically decides whether to use one of three sorting algorithms—InsertionSort, MergeSort, or QuickSort—performs faster than any of these algorithms alone. Their heuristic is a simple stepwise function, much like our χ. Again, neither of these heuristics entails directly computing with anything like the MDP representation we have been using to model agents "from the outside", so to speak. One consequence of this is that such strategies do not transfer to any other metareasoning problems; they are fixed heuristics that work only for the specific problem domain, such as deciding the number of samples or deciding which sorting algorithm to use. Even if such an agent were given useful information, e.g., about the relative reliability or efficiency of one of its computational strategies, it would not know how to use it.

However, we can imagine agents who employ more sophisticated representations. For example, in the simulating scientist example, we used the MDP to show that the scientist is justified in taking the preliminary steps to determine whether exact inference would be feasible. But there is no reason the scientist could not also have carried out such an analysis to determine whether it is worth taking the extra time to make the screening calculation. There is ample evidence that people do track information about their own reasoning, potentially relevant to the informational decision problem, as we have characterized it (e.g., Nelson and Narens 1990). Our framework is supposed to tell us how that information should be used, if at all.

Researchers in artificial intelligence have long realized that explicit metareasoning—wherein the agent reasons about itself using some model of itself—could be useful (for a general review, see Cox 2005). Indeed, these themes date back to the very beginning of AI (McCarthy, 1959; Minsky, 1965), though this work was not initially concerned explicitly with issues of boundedness. Beginning in the 1990s some practitioners have used frameworks very similar to the informational decision problem sketched here, to help guide *search* in particular. Instead of the more traditional heuristic-based strategies, this work is based on assigning a value to all possible next steps of computation, analogous to the value functions considered above, so that the agent can simply choose the computation of highest expected value. The computational actions in these applications range from search in medical diagnosis (Horvitz et al., 1988), to expanding a node in a game tree (Russell and Wefald, 1991), to optimally sampling the value of a random variable (Hay et al., 2012), all of which can be thought of as search problems.

Making this move requires more on the part of the agent. Much more, in fact. When we take an *external* perspective on an agent's deliberative situation, modeling their space of possible computations using an MDP, the only assumption is that the agent has some distribution over action utilities which evolves over time. This distribution could merely reflect the agent's dispositions to act with minimal internal processing, as long as these dispositions are appropriately affected by informational changes. From the internal perspective, on the other hand, in order to apply the value of information framework as sketched above, the agent must also have beliefs about how its beliefs are likely to change after further thinking. Moreover, these beliefs cannot be arbitrary. Ideally, they should be coherent and they should satisfy principle (R) above. This makes intuitive sense in the simulating scientist case. It is perfectly clear what will happen in every possible outcome of the initial calculation, even though one cannot know before performing the calculation what the outcome will be. The problem is small enough that ensuring (R) holds is not terribly cumbersome.

Typically in such problems the search space itself, the space of possible computations, is given. The goal is also clearly defined, e.g., winning a game. The challenge is actually to analyze the situation into something like an informational decision problem, in particular to find good posterior utility estimates for computations. Performing the whole computation first to observe the changes in utility of the outcome defeats the purpose of the informational decision problem, which is supposed to be used for minimizing and optimizing computing time. The most common way to estimate the value of a computation is to partition the space according to certain easily detectable features and learn estimates for the different computation types through experience (Horvitz et al., 1988; Russell and Wefald, 1991). Such estimates routinely violate principle (R), and sometimes the resulting models do not define proper probability distributions. But if one chooses the features in a smart way, they are typically close enough to improve performance. In the game *Othello*®, Russell and Wefald's search algorithm is thirteen times more efficient than the standard alpha-beta search algorithm (Russell and Wefald, 1991). Indeed, for a bounded agent, it may be more beneficial to maintain a simpler but incorrect, or even incoherent, model (of its own mental or computational processes) than a complex, but accurate model. Much of the work in artificial intelligence goes into investigating different meta-level strategies for how deep in the search tree to go for estimating values of immediate actions (Russell and Wefald, 1991; Hay et al., 2012).

While the algorithms explored in this literature are suggestive, it is unclear how closely the strategies correspond to strategies that human agents should or could follow in similar tasks. Moreover, these models allow only one level of metareasoning, and they typically echo Hacking in assuming the cost of this level can be "rounded off". This leaves out two important sources of additional cost that ought to be taken into account for real agents: the cost of calculating the optimal strategy—this is why "myopic" (Russell and Wefald, 1991) or "semi-myopic" (Hay et al., 2012) methods are used—and just as importantly, the cost of formulating the problem in the first place, a point stressed early on by Simon (1955). The fact that an agent could take these costs into consideration just as well, in effect deliberating about how to deliberate, suggest that we might want include them explicitly in our model of the agent (and in the agent's own model of itself).

Despite these shortcomings, to be partially addressed in the next section, this work provides powerful evidence that a detailed metareasoning architecture, employing explicit models of the agent's own computations, can be successful in concrete applications. These systems literally trade off the costs and benefits of taking further computational actions before settling on a concrete action.

4.4.5 Higher-Order Deliberation and the Return of Regress

There is no reason we should be limited to metareasoning at only one level up. While it may usually be imprudent to go beyond one or two levels—reasoning about how to reason about how to reason, and so on—it is evidently something we can do, and perhaps something we ought to do in some cases. It is here that the threat of regress looms. We cannot simply "round off" the cost of metareasoning if we allow arbitrarily high levels of reasoning about reasoning. Ryle (1949) essentially made this point several years before Savage (1954):

> Intelligently reflecting how to act is, among other things, considering what is pertinent and disregarding what is inappropriate. Must we then say that for the hero's reflections how to act to be intelligent he must first reflect how best to reflect how to act? The endlessness of this implied regress shows that the application of the criterion of appropriateness does not entail the occurrence of a process of considering this criterion. (31)

If intelligent action requires something like deliberation, and if we want to regard deliberating, thinking, reflecting, etc. as particular kinds of actions, subject to some of the same kinds of rational evaluation, then in order to act intelligently we would have to deliberate intelligently, which would require intelligently deliberating about how to deliberate intelligently, and so on, *ad infinitum*.

The remaining challenge is therefore to show how an intelligent agent could potentially engage in arbitrarily higher-order metareasoning, going as high in the hierarchy as it needs, without succumbing to regress. Here is a very general sketch of how this is possible, taking a cue from the sampling-based metareasoning agent from Fig. 4.3. Imagine an agent who runs a constant monitor of its situation, in order to make a very simple binary decision: think or act.[13] Whenever presented with a decision problem—any situation in which an action must be taken—this monitoring function is consulted, and immediately and resolutely leads either to an action or to a deliberative episode. Importantly, this applies not just to the initiation of a decision problem, but also *in medias res* when the agent needs to decide whether to continue deliberating, or to cease reflection and act. A skeleton of the associated Markov Decision Process can be depicted as in Fig. 4.4.

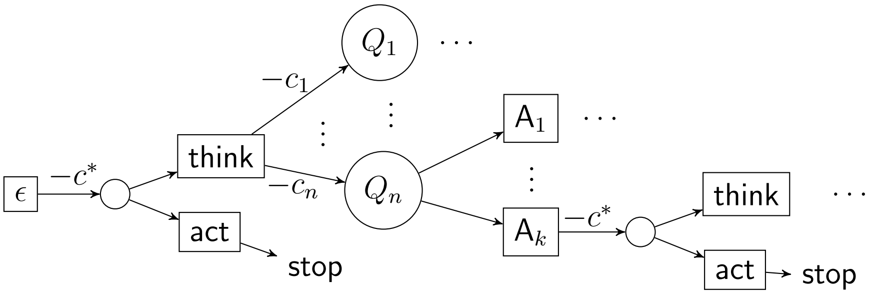

Figure 4.4: Monitoring for higher-order deliberation

The monitoring step is always taken initially, incurring cost c^*, and the effect is either the outcome act, leading directly to stop, in which case a concrete action is taken reflecting the agent's current dispositions, or the outcome think, which leads to further deliberation. In the latter case, the agent has "decided" that it has the luxury of thinking a bit about the problem. At that point, a choice must be made about what course of deliberation to pursue, from among the possible questions $Q_1, \ldots, Q_n, \ldots$ with associated costs $c_1, \ldots, c_n, \ldots$. These could include: sampling from a model of some domain relevant to the first-order problem; asking oneself whether there may be other possible concrete actions that might not yet have been considered, potentially expanding the number of choices to weigh (cf. Douven 2002); assigning

[13] Fodor (1987) refers to the choice between thought and action as "Hamlet's Problem" (140), and claims that this is essentially the frame problem for agents with non-informationally-encapsulated (i.e., not completely modular) processing abilities.

explicit probabilities and utilities to possible outcomes and acts, respectively, so as to carry out explicit calculations (cf. Ch. 2, and Weirich 2004, Ch. 5); or, importantly, asking oneself a higher-level question about how to settle some lower-level deliberative question raised earlier, e.g., about how to answer a question Q_i that has been raised but not answered sufficiently.

Once this deliberative episode is concluded—i.e., after the question is answered, which may simply be to the effect, "I am not sure"—the monitoring step is automatically taken again. This, in turn, may lead to further deliberation or to immediate action, where the action dispositions now reflect the outcome of the previous deliberative step.

There are many variables to specify in this basic schema: how the potential questions are generated, how these questions are answered, and how the utilities depend on answers to questions. I will not try to specify these variables here. The most important detail from the perspective of regress worries is how the basic think/act choice is made, and how costly this step might be, i.e., how high c^* is relative to other costs and utilities. The choice itself is binary, and it seems reasonable to suppose it could reliably depend on only a few easily discernible variables. Something along the lines of χ would be a very simple example, where the decision whether to deliberate further or act immediately is an elementary function of the relative magnitude of the problem. Perhaps the decision should also depend on information about the time course of the decision. Even very rough estimates of these quantities ought to be sufficient. More sophisticated procedures related to *optimal stopping policies* (DeGroot, 2004, Ch. 12-14), including cheap estimations thereof, may also provide appropriate methods of sufficiently low cost. Related ideas about *satisficing* are aimed at providing a cheap but reasonable stopping rule (Simon, 1956). There have furthermore been suggestions, dating back to Wiener (1948) and even Dewey (1922), that immediate action is the *default*, and reflection or deliberation is only the result of a realization ("alarm bells") that relying on current dispositions will not be adequate to the task. Indeed, there are many conceivable ways this simple binary choice might be made, with relatively simple cues.

At any rate, it seems a reasonable agent architecture might well employ such a monitoring function, so long as the costs are low enough to be "rounded off" legitimately. In line with the quotation from Ryle, applying this function does not amount to anything we would want to call "consideration" or "deliberation"; it might be more like a "reflex" involving simple look-up-table-like computations. It ought to be reflexive rather than reflective, as we might say. This is how we thwart regress (cf. Dennett 1981a on ever simpler and dumber homunculi).

To be sure, an agent can ask itself *whether* to deliberate about a problem, thereby considering this question in a more robust, agentive sense. But once that step has been taken, we are already at one of the first Q_i nodes in Fig. 4.4—the decision to pause and think has already been made. At best, asking this question can affect whether an action is immediately taken at the *next* step.

The proposal sketched here is not the only conceivable way of incorporating higher-order deliberation into an agent architecture, but I would like to offer it as a plausible schema for understanding how people determine whether or not to stop and think before acting. If this is right, there are interesting empirical details about what cues go into determining this. Anecdotal evidence suggests people are reasonably good at knowing when to act immediately, e.g., as in the highway exit example, or to stop and think before acting.[14] My suggestion is that this basic ability also supports the capacity to engage in higher-order deliberation. Assuming the existence of a simple reflex-like mechanism for determining when to end deliberation, our agent can step up the metareasoning hierarchy as high as it likes. I hypothesize that such an architecture will perform well in terms of the bounded rationality framework sketched in the previous chapter. Testing the relative fitness of such agents, through simulations or otherwise, is a venture I leave for future work.

[14] They are of course not perfect, witness the complementary popular advice "Think before you speak" and "Try not to over-think things".

4.5 Using Informational Decision Theory

Let us take stock. Savage's challenge was to integrate the cost of thinking into decision theory in a way that allows trading off the value of more information with the cost of obtaining, accessing, or generating that information. We have seen that this is not only possible, but can be quite fruitful in understanding aspects of intelligent behavior, as Savage suggested it would be. Moreover, the framework underlying the informational decision problem is just an application of classical decision theory. If we are modeling an agent "from the outside", the transitions in the corresponding MDP simply represent possible mental or computational transitions the agent could take, and we can determine what the best path for that agent would be, using our own estimate of the statistics of the world and our own understanding of the agent's computational abilities. This is essentially what we were doing with boundedly rational analysis in the previous chapter. If we want an agent to be able to formulate its own first-order decision problem, including costs of computation, as an informational decision problem and solve for an optimal policy at a meta-level before deciding which first-order deliberative path to pursue, this is possible too, provided the agent has a means of obtaining estimates of the costs and benefits of possible computational steps. So does this mean Savage's challenge has been met? In this section, I will discuss the sense in which I think the challenge has been met, though I will also raise some puzzles and mention some aspects of the project that I believe merit more work and exploration.

Standard decision theory could be thought of as universal in the following sense. Given an agent's preferences and beliefs, the theory provides a determinate answer to the question of what to do: take any action whose expected utility is not dominated. So long as the situation satisfies the assumptions of the theory—e.g., well defined probabilities and utilities—there is no ambiguity in what actions the theory says are admissible. The problem we started with was that deliberation and cost of thinking are usually left out of the equation. Now that we know how to incorporate them into the theory, is it still universal in the relevant sense? That is, given an agent's preferences and beliefs, and given the agent's computational limitations, does the theory always provide a determinate answer of what to do?

Yes and no. To see the difficulty in giving a straight answer, suppose we have before us an MDP accurately representing an agent's situation, including all the possible steps in thought the agent could take before choosing some action. As long as there is at least one optimal strategy σ for such an MDP, the theory says the agent should adopt such a strategy. In some sense σ captures the best way the agent could use its computational resources to choose the best action. After all, formally speaking, the informational decision problem is just a standard sequential decision problem. So we might say the theory is still universal in the relevant sense. However, suppose now our agent must actually follow some strategy, but it does not know σ defines the best strategy. For example, suppose while the agent is capable of following σ, determining that σ is optimal is exceedingly difficult. What is our agent to do?[15]

There are at least two possible responses. One is to insist that the thing to do is act according to σ, whether or not the agent knows it. By assumption we have included every possible sequence of events up to the time of action, and surely, if there is any path at all that our agent ought to follow, it lives somewhere in this model. If the agent should not opt for the best strategy, then which one should it take?

Another response is to reject the very idea that there could be such a situation. When we say that the MDP represents all of the sequences of thought that the agent could possibly follow, we do not just mean that the agent is psychologically capable of undergoing these transitions. We must have something more agentive in mind, in order for the theory to apply to actual deliberating agents. The MDP must only include paths that the agent *knows how* to follow. Compare the situation to a person attempting a maze. There is an obvious sense in which the person has the ability to traverse the maze on the first attempt, and there is a sense in which that is what the person should do. But, as in the case of Savage's bet on a remote digit of π, this recommendation is wholly unhelpful to the person who does not know which path leads to the goal. A better recommendation might be something like: Always try the leftmost path first and backtrack when necessary, or some other clearly executable strategy. If decision theory is going to be of any use in actual decision making, it

[15] See (Nozick, 1981, 300) for a similar worry about applying decision theory to bounded agents.

should make recommendations that can actually be followed.

I believe there is a grain of truth in both responses. But they each lead to absurdities when taken too far. The first makes it seem as though our theory is committed to a rejection of the "ought implies can" principle; the second, when brought to its extreme, makes it seem as though there is only one thing an agent can do, namely what it does. The way to resolve this apparent puzzle is to distinguish three separate uses of the informational decision problem framework: in designing intelligent agents, in assessing agents as more or less rational, and in deliberating about what to do.[16] These three are of course closely connected, but they relate to the mathematical theory in distinctive ways. While I think we can answer Savage's challenge with a mostly positive answer, this inevitable trifurcation makes for a slightly less unified theory than one may have expected or hoped, since the components of the model may be interpreted in different ways in the three contexts.

That said, this is a natural, and I believe welcome, consequence of incorporating computational costs. It is significant that the informational decision theory framework allows these distinctions to be made in the first place. Classical decision theory, as typically presented, has no resources to talk about what one is cognitively able to do. We could take the same hard line and insist that there may be paths of deliberation a person ought to take even when that person lacks the cognitive know-how. These cases arise in the context of assessment, as I will explain shortly. But there are also cases in which the model should include only deliberative paths the agent knows how to follow. This is especially appropriate when one wants to use this kind of framework in actual deliberative situations.

4.5.1 Design

The problem of designing an optimal agent is not well defined in general. At best one can optimize some aspect of an agent, holding all other aspects fixed, relative to some goals or utilities arising from the agent's action possibilities. If the task in question is sufficiently small and manageable, it may be possible to write down an MDP

[16] There is a fourth use of this framework, as part of a more general boundedly rational analysis methodology for *discovering* how minds in fact work. This was our topic in Chapter 3.

representing all feasible computation paths. Provided we have some utility measure on outcomes, we could then solve the MDP for the optimal strategy and program the agent accordingly. In some simple cases, this is possible (e.g., Russell and Subramanian 1995). In more complicated domains this would be impractical. Nonetheless, at a very high level of abstraction, this does capture what agent programmers are aiming to do. The probability distributions then represent the designer's own estimates about what state the world is likely to be in when the agent is in a given internal state, and the utilities ultimately derive from what we want the agent to accomplish. When we say, pointing to the MDP, that an agent *could* perform action A, or *could have* performed A when in fact it performed B, it may simply mean that we could rewrite the agent's program so that it performs A in any relevantly similar situation. The informational decision problem, or the corresponding MDP, is therefore used as a guide, and as a means for generating predictions about what a given agent will do in its environment, as the designer predicts it to be. This is all true whether we are talking about concrete external actions or internal computational actions.

The design question becomes most interesting, however, when we are trying to design agents who reason about their own reasoning. For agents with complex reasoning abilities, it is not at all clear what the ideal strategy would look like. Anecdotal evidence from the work presented and reviewed above, plus the observation that human metareasoning often seems beneficial, suggest that an agent who uses metareasoning wisely may be better off than one who uses no metareasoning. But this leaves open many unsolved questions. For instance, how detailed should the agent's model of itself be? Under what conditions should an agent take a step up in the metareasoning hierarchy? At what point should an agent give up on metareasoning and just reason, or act? These are all practical questions and will presumably depend on the application and circumstance, as well as details about the agent's architecture.

4.5.2 Assessment

From the design point of view, our question about what the agent should do when it does not know what is optimal does not have much bite. The agent just does what we program it to do, or perhaps it malfunctions. All "oughts" derive from the designer's goals and take on an operational meaning. If it will not take A unless it "knows" to take A, then we must ensure it knows to take A.

When assessing a person's deliberative strategy as more or less rational, the question becomes more pressing. If we judge that they could have followed A but followed B instead, and A would have been better by that person's own lights, we are compelled to say the person was acting irrationally. But here we do have to be careful about what we say the agent could do. Again, consider the analogy to the maze explorer. Would we really want to criticize someone for not being able to find the goal on the first attempt?

On the other hand, suppose by sheer luck the person does proceed through a path straight to the goal on the first attempt. In that case we may not want to say, in retrospect, the person was foolish to follow it, that it would have been wiser to take a more prudent approach. Much less do we want to say that this path was impossible. Thus we are back to the original dilemma: which paths should be included among those that the agent is intuitively able to follow? The dilemma involves some old and difficult issues—about rational assessment, ability, knowledge and know-how—which are not specific to metareasoning. I have nothing new to say about those here. But there is a pragmatic way of approaching the dilemma which I believe is promising.[17]

One good thing about the practice of rational assessment is that, in various ways, both socially and individually, it can change and shape future behavior. By noticing, appreciating, and praising good reasoning, and criticizing bad reasoning, we can become better reasoners. This is so even if in the episode the person, strictly speaking, could not have done differently. We do not need to put ourselves in the shoes of the deliberating agent for this practice to be successful. We can simply acknowledge that under such circumstances, some or other sequence of reasoning step would have led to the best outcome. Importantly, this can happen outside the heat of deliberation, where cool, sober reflection reveals what one ought to have done.

In as far as rational assessment enhances an agent's overall rationality, it is also pertinent to the design problem. When the agent follows process B instead of A

[17] Similar approaches to rational assessment appear, e.g., in Bratman (1987) and Stich (1990).

where it should have followed A, it would obviously be preferable if the agent could modify its own program so that it follows A next time. This kind of learning problem is extremely well studied in AI and neighboring fields (see, e.g., Sutton and Barto 1998), though it remains to be seen how close one can get to something that looks like learning from high-level, sophisticated rational assessment. I see this as an important open problem of bringing together work on statistical and machine learning with philosophical work on practical reasoning and action theory. How can these two kinds of learning fit into a single agent architecture?

To summarize, we can use informational decision theory to assess an agent's deliberative strategy, and there may be good reason to say of that agent that a certain strategy should have been pursued, even when the agent was not explicitly aware that this strategy would have been preferable. In the context of assessment, "ought to have" need not imply "could have" in any strong sense, as long as "ought to have" does allow for the possibility of "will be able to".

4.5.3 Deliberation

One peculiarity of deliberation and decision making is that sometimes the best way of making a decision is by not deliberating at all. So much is clear in the case of the driver quickly approaching the fork in the road.[18] Indeed, in the theoretical framework presented here, it is easy to write down an MDP in which the optimal strategy is simply to act at the first choice point. In these cases, the thing to do, according to informational decision theory, is not to appeal to any theory. This alone shows that informational decision theory, even with all of the costs of computation taken into account, cannot be applied universally in the context of deliberation. We do not even need a regress argument to show this.

This is not to say the informational decision problem framework is irrelevant

[18] There are of course many other cases. As Skyrms points out, this may be rational not only because of time and cost pressures, but because deliberation may lead to unstable oscillation and indecision (Skyrms, 1990, pp. 102-103). In the context of planning, Bratman (1987) has also argued forcefully for the possibility that deliberation, specifically in the form of intention reconsideration, ought sometimes to be avoided. Making sense of this in the framework described here is a project I hope to pursue in later work.

to actual decision making. On the contrary. The original work on the value of experiments by Raiffa and Schlaifer (1961) and others was intended for use in actual decision situations, where time is spent on solving some version of the informational decision problem before making an important large-scale decision. We also gave an example, where the simulating scientist takes some time out of the first-order task to assess at the second order what first-order strategy to take. The value-of-information story tells us which lines of inquiry an agent ought to take. Certainly, if the agent can make use of that information—even approximately, as with the agents who "reason" about sampling and sorting—it could stand to gain.

One thing these apparently unproblematic cases have in common is that the size of the problem is small, and the relative cost of carrying out this extra meta-computation is insignificant compared to the costs involved in the first-order decision problem. In Hacking's words, "the cost of meta-thinking gets rounded off". From the outside, we can show using an MDP formalizing the agent's deliberative situation that this is true in a very precise sense. But, crucially, this will not always be the case. How can an agent know, in the heat of deliberation, when it can and when it cannot? The proposal I made in §4.4.5 is that this binary choice can be much simpler than the more general metareasoning problem as codified by the formal framework in §4.3.1, enough so that its costs are indeed negligible. There are conceivable agent architectures—which I believe human agents exemplify—that make use of very simple monitoring mechanisms whose costs are minimal: they can be rounded off by us the theorists, and not even considered explicitly by the agent. They run constantly, so there is no decision to be made about whether to monitor. As long as there are reliable cues to which the agent can be adequately attuned, leading to thought or action as appropriate, the only remaining problem is *what* to think about whenever the decision to think at all has been made.

Again, for this latter question, there is no reason to assume an agent should not use some version of the value of information framework to assess its own (meta)reasoning possibilities. Of course, this is only useful insofar as the paths considered and assessed are paths the agent *knows how* to follow, or at least knows how to figure out how to follow with perhaps further thought and deliberation. From the deliberative standpoint, unlike from the standpoint of assessment (which may nevertheless follow deliberation, and thus help in the next episode of deliberation), it is not helpful to formulate the deliberative situation using computational actions one does not know how to perform. For instance, in deciding how to decide how to bet in Savage's example, it would not be helpful for the agent to consider the strategy of computing the nth digit of π, unless it has some way of actually computing it. If it does, a more concrete representation of what it needs to do to compute it, how much it will cost, and how likely it is to be correct, all ought to be taken into account.

Thus we see that the interpretation of the informational decision problem may be slightly different, depending on whether we are using it to design an agent, to assess an agent, or to guide our own deliberation.

4.6 Conclusion

Why did Savage suspect that incorporating the cost of computation or calculation into decision theory would entail paradox or regress? If rational deliberation always requires a separate act of rationally deciding how to proceed in that deliberation, then rational deliberation could never begin. In line with the bounded rationality framework adopted in this dissertation, there is no reason to accept this assumption. As we saw in Chapter 3, sometimes inflexible, non-deliberative, reflex-like or look-up-table strategies are more boundedly rational.[19] Having given up the problematic assumption, Savage's primary challenge remains: to show how we can assess an agent in deliberation, weighing costs and benefits of that deliberation, and in a way that could be useful to an agent already actively engaged in deliberation.

I have proposed—for agents capable of a very modest amount of self-reflection—to study metareasoning and deliberation using standard tools for sequential decisions, namely Markov Decision Processes capturing the value of information accorded by different computational or deliberative actions and strategies. I suggested construing

[19] In an amusing twist, it may be that the arguments for "fast and frugal heuristics" (e.g., Gigerenzer and Goldstein 1996) apply most clearly to metareasoning problems, whose purpose is to regulate our more complex cognitive abilities that require more than simple heuristics.

the thought process as sequences of questions and answers to oneself, where different lines of questioning would vary in their expected usefulness for improving action selection. This allows us as theorists or observers to assess, as well as design, agents and their deliberative activities. It also has the potential for aiding in the context of deliberation, so long as it is both possible and boundedly rational for the agent to use these tools (whether implicitly or explicitly) in a given context.

It is always possible to "elevate" to the next level, to deliberate about how to deliberate, and sometimes this may be useful. But it does raise the question of how an agent could have this flexibility, without reinviting regress worries. My proposal is that every step of deliberation is preceded by a reflex-like monitoring step, determining whether the agent will act immediately under its current disposition, or stop to think about what to do. If this monitoring step is cheap enough, it can run constantly so that there is never any reason to factor its costs into one's deliberation. Such an architecture strikes me as very likely to be boundedly rational in the requisite sense, and I believe we in fact employ such a monitoring mechanism.

Of course, in the context of deliberation, once one has decided to deliberate, there is the question of how best to spend that time deliberating. The picture we get from the value of information account offered here tells us abstractly about the logic of this activity, which, as we have seen, can be helpful in understanding the general structure of good reasoning. To ask for anything more concrete is just to ask how one ought to reason in a particular case.

Chapter 5

Conclusion

My aim in this dissertation has been to tell a story about the nature of human inference; to encapsulate and explore what I see as an emerging picture of how inference works, across several fields of inquiry; and to suggest new problems, ideas, and methodological considerations for moving forward. Specifically, I first presented a general view of the mind as a sampling engine, which I claimed offers a novel and compelling way of thinking about subjective probability and how inference could be probabilistically grounded. I then discussed this view in the context of the Bayesian program in cognitive psychology. The sampling view helps to address some methodological and empirical challenges that have been raised for this program. I proposed a thoroughgoing *boundedly rational analysis*, according to which sampling algorithms appear to be natural candidates for concrete cognitive mechanisms. Once the challenges to the Bayesian program are answered, this tradition can be seen as a fruitful source of insights into the nature of inference, and indeed of the sampling view of inference in particular, allowing the exploration of richer representational schemes and computations over those richer representations.

With bounded resources and rich representational capacity, a very natural topic—which has nonetheless not yet been explored in this literature—is *metareasoning*, or reasoning about one's own reasoning, mental state, resources, or decision problem. I showed how the introduction of metareasoning capabilities makes the boundedly rational analysis strategy more complicated, but also potentially more interesting.

I believe a deeper understanding of the nature of metareasoning is critical to understanding the nature of ordinary inference, and indeed of cognition more broadly. Following I. J. Good, I proposed an analysis of metareasoning in terms of the *value of information*, turning this framework "inside the head", where information-producing actions come from internal computation rather than external observation. I also explored how metareasoning could function in human (and perhaps in artificial) agents in such a way that arbitrarily higher-order reasoning about reasoning could be effected without succumbing to vicious regress.

In the introductory chapter, I framed the dissertation in terms of the *why?* of inference: understanding why we find ourselves making certain inferential transitions rather than others. What have we learned about this question? Human inference is shaped both by the "hardware" and resources available for carrying out the inference, and the character of the problem(s) being solved at the time, that is, what project(s) the agent is engaged in when the inference is made. The picture of inference drawn in the first two thirds of this dissertation is an attempt to delineate much of the *how?* of inference, by appealing both to what we know of brains work, or could work, and what we can say about what an ideal inference engine would look like, subject to general resource limitations and general architectural constraints that we know are in place. Rational analysis, and its bounded variant, offers an answer to *why?* questions generally, by specifying what problem the mind is assumed to be solving. In the case of boundedly rational analysis of inference specifically, the basic aim is to call to mind information—presumably information whose plausibility is well supported by evidential and other considerations—that has the best chance of being relevant to acting in a way conducive to the agent's assumed goals. Inference works appropriately when it is fulfilling this aim. There is empirical evidence suggesting that human inference is in many ways optimized, particularly once we take bounds into account. I have also argued that, often, the intelligent guidance of inference is based on (typically unconscious) metareasoning and monitoring mechanisms, that is, mental mechanisms whose role is to reason about other internal reasoning processes. In some cases, the nature of the problem is sufficiently difficult that metareasoning becomes indispensable. In addition to being a story about *how* inference works, this kind of analysis seeks to explain *why* our inferential mechanisms work as they do. Boundedness and metareasoning are part of that rational explanation.

As remarked in the introductory chapter, this is only the beginning of a thorough study of the topics explored and introduced here. A number of topics for further exploration have already been identified, including suggestions for future experiments (Appendix A.2, 4.4.5, *inter alia*), mathematical questions about classes of sampling and other algorithms (e.g., 3.5.3), and philosophical questions about the role of decision theory, broadly conceived, in the analysis of human decision making (4.5). In addition to these directions, I see the most promising and important development of these ideas in their further incorporation and unification with other methods, ideas, examples, and concepts familiar from logic and philosophy. In fact, I view much of the work done here as laying some of the groundwork and foundations for such an integration. From the perspective of logic and philosophy, I hope to have illustrated some of the consequences of taking a psychological and computational perspective seriously, with a view toward realism and implementation. In the other direction, the rich traditions from logic and philosophy have carefully mapped out and analyzed so many fascinating and subtle phenomena—ranging from natural language semantics to planning and agency—which have hardly been broached in this thesis.

Some topics that I myself have worked on, including natural language inference and "natural logic" (Icard, 2012; Icard and Moss, 2014) and the problem of intention revision (Icard et al., 2010), to take just two examples, have been directly motivated by considerations concerning resource boundedness. In fact, these were the sorts of problems that motivated this dissertation project to begin with. These domains, as well as countless others from the philosophical and logical literatures, promise challenging and potentially rewarding examples to try to subsume under the framework presented here, going beyond toy examples and instances already familiar from the cognitive psychology literature. In this way, cognitive science has much to gain. In other direction, one often sees discussion of resource limitations as motivation for various philosophical claims. Couching those claims in the context of a working computational model of the mind, and exploring the consequences through simulation or related types of analysis, promises a concrete and transparent perspective.

Appendix A

A.1 Boltzmann Machine as Gibbs Sampler

In this Appendix we explain the sense in which the activation rule for the Boltzmann Machine carries out Gibbs sampling on an underlying Markov random field. This fact is folklore in cognitive science, but there does not seem to be a clear presentation of the fact in print.

Gibbs sampling is an instance of the Metropolis-Hastings algorithm, which in turn is an example of Markov chain Monte Carlo (MacKay, 2003). Suppose we have a multivariate distribution $P(X_1, \ldots, X_n)$ and we want to draw samples from it. The Gibbs sampling algorithm is as follows:

1. Specify some initial values for all the random variables $\mathbf{y}^{(0)} = (y_1^{(0)}, \ldots, y_n^{(0)})$.

2. Given $\mathbf{y}^{(r)}$, randomly choose a number $i \in \{0, \ldots, n\}$, and let $\mathbf{y}^{(r+1)}$ be exactly like $\mathbf{y}^{(r)}$, except that $y_i^{(r+1)}$ is redrawn from the conditional distribution $P(X_i \mid y_{-i})$.

3. At some stage R, return some subset of $\{\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(R)}\}$ as samples.

Note that the sequence of value vectors $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(r)}, \ldots$ forms a Markov chain because the next sample $\mathbf{y}^{(r+1)}$ only depends on the previous sample $\mathbf{y}^{(r)}$. Let $q(\mathbf{y} \to \mathbf{y}')$ be the probability of moving from $\mathbf{y}$ at a given stage r to $\mathbf{y}'$ at stage $r + 1$ (which is the same for all r, and 0 when $\mathbf{y}$ and $\mathbf{y}'$ differ by more than one coordinate.). And let $\pi^r(\mathbf{y})$ be the probability of being in state $\mathbf{y}$ at stage r. Then we clearly have:

$$\pi^{r+1}(\mathbf{y}') = \sum_{\mathbf{y}} \pi^r(\mathbf{y}) q(\mathbf{y} \to \mathbf{y}') \ .$$

By general facts about Markov chains, it can be shown that this processes reaches a unique stationary distribution, i.e., $\pi^r = \pi^{r+1}$. This is the unique distribution π^* for which the following holds:

$$\pi^*(\mathbf{y}') = \sum_{\mathbf{y}} \pi^*(\mathbf{y}) q(\mathbf{y} \to \mathbf{y}') \ .$$

Since this equation also holds for P, that shows the Markov chain converges to $P = \pi^*$.

Recall the Boltzmann Machine is defined by a set of nodes and weights between those nodes.[1] If we think of the nodes as binary random variables $(X_1, \ldots, X_n)$, taking on values $\{0, 1\}$, then the weight matrix $\mathbf{W}$ gives us a natural distribution $P(X_1, \ldots, X_n)$ on state space $\{0, 1\}^n$ as follows. First, define an *energy function* E on the state space:

$$E(\mathbf{y}) = -\frac{1}{2} \sum_{i,j} W_{i,j} y_i y_j \ .$$

Then the natural *Boltzmann distribution* is given by:

$$P(\mathbf{y}) = \frac{e^{-E(\mathbf{y})}}{\sum_{\mathbf{y}'} e^{-E(\mathbf{y}')}} \ .$$

Note that the denominator becomes a very large sum very quickly. Recall from our earlier discussion that the size of the state space is 32 with 5 nodes, but over a trillion with 40 nodes. Suppose we apply the Gibbs sampling algorithm to this distribution. Then at a given stage, we randomly choose a node to update, and the new value is determined by the conditional probability as above. This step is equivalent to applying the activation function for the Boltzmann Machine (where in the following,

[1] We ignore bias terms here to simplify the presentation. See Rumelhart et al. (1986) for the general formulation.

$\mathbf{y}'$ is just like $\mathbf{y}$, except that $y'_i = 1$ and $y_i = 0$):

$$\begin{aligned}
P(X_i = 1 \mid \{y_j\}_{j \neq i}) &= P(X_i = 1 \mid \mathbf{y} \text{ or } \mathbf{y}') \\
&= \frac{P(\mathbf{y}')}{P(\mathbf{y} \text{ or } \mathbf{y}')} \\
&= \frac{e^{-E(\mathbf{y}')}}{e^{-E(\mathbf{y})} + e^{-E(\mathbf{y}')}} \\
&= \frac{1}{1 + e^{E(\mathbf{y}') - E(\mathbf{y})}} \\
&= \frac{1}{1 + e^{-\sum_j W_{i,j} y_j}} \\
&= \frac{1}{1 + e^{-net_i}} \; .
\end{aligned}$$

Thus, the above argument shows that the Boltzmann Machine (eventually) samples from the associated Boltzmann distribution.

Incidentally, we can also see now why net_i is equivalent to the log odds ratio under the Boltzmann distribution (recall the discussion in §2.8 of the Yang and Shadlen 2007 experiment):

$$\begin{aligned}
\frac{P(\mathbf{y}')}{P(\mathbf{y})} &= \frac{e^{-E(\mathbf{y}')}}{e^{-E(\mathbf{y})}} \\
&= e^{E(\mathbf{y}) - E(\mathbf{y}')} \\
&= e^{\sum_j W_{i,j} y_j} \\
&= e^{net_i} \; .
\end{aligned}$$

And thus,

$$\begin{aligned}
\log \frac{P(\mathbf{y}')}{P(\mathbf{y})} &= \log\left(e^{net_i}\right) \\
&= net_i \; .
\end{aligned}$$

A.2 Softmax versus Sampling Rule

In this Appendix, we offer some observations about the relation between the softmax (or generalized Luce-Shepard) choice rule and the sampling-based decision rules discussed in this chapter. Suppose we associate with a subject a probability function $P(\cdot)$ on sample space $\mathcal{H} = \{H_1, \ldots, H_n\}$ and a utility function u over actions $\mathcal{A} = \{A_1, \ldots, A_m\}$.

The softmax rule says that the subject will give response A with probability

$$\frac{e^{v(A)/\beta}}{\sum_{A' \in \mathcal{A}} e^{v(A')/\beta}},$$

where v is some value function, which in this case we will assume is the log expected utility:

$$v(H) = \log \sum_{H \in \mathcal{H}} P(H)u(A, H).$$

As a representative example of a sampling based rule, recall Decision Rule B:

> Decision Rule B: Suppose we are given a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values. To select an action, take R samples, $H^{(1)}, \ldots, H^{(R)}$, using $\mathcal{M}$, and let best be the set of actions that receive the largest summed utilities, i.e.,
>
> $$\text{BEST} = \{A_j : \sum_{i=1}^{R} u(A_j, H^{(i)}) \text{ is maximal}\}.$$
>
> Take action $A_j \in \text{BEST}$ with probability $\frac{1}{|\text{BEST}|}$.

In the specific case of a certain kind of estimation problem—where $\mathcal{H} = \mathcal{A}$ and the utility of an estimate is 1 if correct, 0 otherwise; thus expected utility and probability coincide—it is easy to see that the softmax rule with $\beta = 1$ and Decision Rule B with $R = 1$ (or $R = 2$) are equivalent. The probability of returning hypothesis H is just $P(H)$, i.e., we have probability matching.

Unfortunately, even restricting to these simple estimation problems, the relation between the two rules as functions of β and R is intractable and varies with the probability distribution $P(\cdot)$, as Vul (2010) points out. Thus, beyond $\beta = R = 1$ it is hard to study their relationship. As mentioned in the text, both can be used to fit much of the psychological data, though one might suspect the sampling rule is more reasonable on the basis of computational considerations.

Interestingly, for more general classes of decision problems, these two classes of rules can be qualitatively distinguished. The Luce Choice Axiom, from which the Luce choice rule was originally derived (Luce, 1959), gives a hint of how we might do this. Where $P_S(T)$ is the probability of choosing an action from $T \subseteq \mathcal{A}$ from among options in $S \subseteq \mathcal{A}$, the choice axiom states that for all R such that $T \subseteq R \subseteq S$, we have:

$$P_S(T) \;=\; P_S(R)\; P_R(T)\ .$$

It is easy to see the softmax rule satisfies the choice axiom for all values of β:

$$\begin{aligned} P_S^{\text{softmax}}(T) &= \frac{\sum_{A\in T} e^{v(A)/\beta}}{\sum_{A\in S} e^{v(A)/\beta}} \\ &= \frac{\sum_{A\in R} e^{v(A)/\beta}}{\sum_{A\in S} e^{v(A)/\beta}} \cdot \frac{\sum_{A\in T} e^{v(A)/\beta}}{\sum_{A\in R} e^{v(A)/\beta}} \\ &= P_S^{\text{softmax}}(R)\; P_R^{\text{softmax}}(T)\ . \end{aligned}$$

For estimation problems, Decision Rule B with $R = 1$ also satisfies this axiom. However, in more general contexts, even for the case of $R = 1$, it does not. Perhaps the simplest illustration of this is the decision problem in Fig. A.1, where $\mathcal{H} = \{H_1, H_2\}$ and $\mathcal{A} = \{A_1, A_2, A_3\}$, and $\epsilon > 0$.

	H_1	H_2
A_1	$1-\epsilon$	$1-\epsilon$
A_2	2	0
A_3	0	2

Table A.1: Distinguishing softmax and sample-based decision rules

We clearly have $P^{\textsc{Rule B}}_{\{A_1,A_2,A_3\}}(\{A_1\}) = 0$. Yet, as long as $P(H_1), P(H_2) > 0$, we have

$$P^{\textsc{Rule B}}_{\{A_1,A_2,A_3\}}(\{A_1, A_2\})\; P^{\textsc{Rule B}}_{\{A_1,A_2\}}(\{A_1\}) > 0 \; ,$$

showing the violation of the choice axiom. As the softmax rule satisfies the axiom, this gives us an instance where we would expect different behavior, depending on which rule (if either) a subject is using. When presented with such a problem, a softmax-based agent would sometimes choose action A_1. In fact, the probability of choosing A_1 can be nearly as high as $^1/_3$ (for example, when $\beta = 1$ and ϵ is very small) if H_1 and H_2 are equiprobable. However, for any set of samples of any length R, one of A_2 or A_3 would always look preferable for a sampling agent. Thus, such an agent would never choose A_1. It would be interesting to test this difference experimentally.

Appendix B

B.1 Anderson's Rational Model of Categorization

Anderon's (1990; 1991a) rational model of categorization defines a prior distribution $P(Z_n)$ for clusters, and likelihoods for features given clusters $P(Y_k|Z_n)$, $P(X_k|Z_n)$—features are assumed to be independent, conditional on the cluster—so that the probability of a clustering can be inferred using Bayes Rule (repeated from Eq. (3.2)):

$$P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;=\; \frac{P(\mathbf{X}_N, \mathbf{Y}_{N-1} \mid Z_n)\; P(Z_n)}{\sum_{Z'_n} P(\mathbf{X}_N, \mathbf{Y}_{N-1} \mid Z'_n)\; P(Z'_n)}\;.$$

The likelihood for a feature given a clustering is given by a beta distribution:

$$P(Y_k = v|Z_n) = \frac{\#_v + \beta}{\# + 2\beta}\;,$$

where # is the number of objects Z_n clusters together with the kth object; and $\#_v$ is the number of those objects in the same cluster that have v as their Y-value. In all of the simulations we ran, $\beta = 1$, so we are effectively using the Laplace prior. As n increases, the likelihood converges to the empirical frequency $\#_v/\#$.

The prior term for $P(Z_n)$ has one free parameter c, the *coupling parameter*, which determines how likely an object is to belong to a brand new clustering. The explicit form of the prior is rather complicated (see Anderson 1990, 1991a or Sanborn et al. 2010); it is easiest to understand as resulting from a sequential process so that clustering Z_{n+1} extends Z_n with probabilities as follows:

$$P(Z_{n+1} = j|Z_n) = \begin{cases} \frac{c \cdot M_j}{(1-c)+c \cdot n} & \text{if } j \text{ assigns the new object to an old cluster} \\ \frac{1-c}{(1-c)+c \cdot n} & \text{if } j \text{ assigns the new object to a new cluster} \end{cases},$$

where M_j is the number of objects already in the cluster that j assigns the new object, according to Z_n. With this prior, the more often objects are categorized as coming from a particular cluster, the more likely new objects are to fall under that cluster, which is why this is often called a “rich get richer” scheme.

Appendix C

C.1 Informational Decision Problems as Markov Decision Processes

Here we demonstrate in what sense an informational decision problem can be viewed as a Markov Decision Process. This is more or less folklore. See, e.g., Bernardo and Smith (1994) or Hay and Russell (2011). See Sutton and Barto (1998) for a textbook treatment of MDPs.

First, let us define the probability $p_{\mathbf{E}}$ of following sequence $\mathbf{E}$ given a strategy σ. Define by induction on the length of $\mathbf{E} \in \mathcal{E}^*$ the probability $q_{\mathbf{E}}$ of reaching sequence $\mathbf{E}$ under strategy σ:

$q_E = \mathbb{1}_{\sigma(\epsilon)=E}$ (indicator function giving 1 if $\sigma(\epsilon) = E$, 0 otherwise).

$q_{\mathbf{E}E} = q_{\mathbf{E}} \cdot \sum_{\mathbf{o} \in \mathcal{O}(\mathbf{E})} [P(\mathbf{o}) \cdot \mathbb{1}_{\sigma(\mathbf{o})=E}]$.

The probability $p_{\mathbf{E}}$ of following sequence $\mathbf{E}$ and then stopping is:

$$p_{\mathbf{E}} = q_{\mathbf{E}} \cdot \sum_{\mathbf{o} \in \mathcal{O}(\mathbf{E})} [P(\mathbf{o}) \cdot \mathbb{1}_{\sigma(\mathbf{o})=\mathsf{stop}}].$$

This definition can be generalized to the probability $p^{\mathbf{o}}_{\mathbf{E}}$ of sequence $\mathbf{E}$, starting from observational state $\mathbf{o}$, simply by redefining q_E as $q^{\mathbf{o}}_E = \mathbb{1}_{\sigma(\mathbf{o})=E}$. We can similarly generalize the value function on sequences to start from any observational state $\mathbf{o}$:

$$u_{\mathbf{o}}(\mathbf{E}) = \sum_{\mathbf{o}' \in \mathcal{O}(\mathbf{E})} P(\mathbf{o}'|\mathbf{o}) \max_i \mathbb{E}(U_i|\mathbf{o}\mathbf{o}'),$$

so that

$$V_{\mathbf{o}}(\mathbf{E}) = u_{\mathbf{o}}(\mathbf{E}) - c(\mathbf{E}),$$
$$V_{\mathbf{o}}(\sigma) = \sum_{\mathbf{E} \in \mathcal{E}^*} [p^{\mathbf{o}}_{\mathbf{E}} \cdot V_{\mathbf{o}}(\mathbf{E})].$$

This is just the usual function V, but the costs of previous experiments leading to outcome $\mathbf{o}$ are left off. We can then state the following theorem.

Theorem 1. *Given informational decision problem $\mathcal{M}$, we can define an equivalent MDP such that the valuation function for $\mathcal{M}$ satisfies the Bellman equation:*

$$V_s(\sigma) = R(s, \sigma(s)) + \sum_{s' \in S} T(s, \sigma(s), s') V_{s'}(\sigma) \ . \tag{C.1}$$

Proof. Given $\mathcal{M}$ we define the MDP (S, A, T, R) as follows:

* $S = \hat{\mathcal{O}} \cup \{\mathbf{o}_\perp | \mathbf{o} \in \hat{\mathcal{O}}\}$,
  the set of states ;

* $A = \mathcal{E} \cup \{\mathsf{stop}\}$,
  the set of actions ;

* $T(\mathbf{o}, E, \mathbf{o}o) = P(o|\mathbf{o})$
  $T(\mathbf{o}, \mathsf{stop}, \mathbf{o}_\perp) = 1$,
  the transition probabilities ;

* $R(\mathbf{o}, E) = -c(E)$
  $R(\mathbf{o}, \mathsf{stop}) = \max_i \mathbb{E}(U_i|\mathbf{o})$,
  the state rewards.

Provided we define $V_{\mathbf{o}_\perp}(\sigma)$ to be $R(\mathbf{o}, \mathsf{stop})$, we can prove that Equation (C.1) holds.

Suppose $s = \mathbf{o}$ is a reachable state under σ and consider two cases: either $\sigma(\mathbf{o}) = \mathsf{stop}$ or $\sigma(\mathbf{o}) = E$. In the former case,

$$V_{\mathbf{o}}(\sigma) = V_{\mathbf{o}}(\mathsf{stop}) = \max_i \mathbb{E}(U_i|\mathbf{o}) = R(\mathbf{o}, \mathsf{stop}) = R(\mathbf{o}, \sigma(\mathbf{o})).$$

In the latter case,

$$\begin{aligned} V_s(\sigma) &= \sum_{\mathbf{E} \in \mathcal{E}^*} [p^{\mathbf{o}}_{\mathbf{E}} \cdot V_{\mathbf{o}}(\mathbf{E})] \\ &= -c(E) + \sum_{o \in \mathcal{O}(E)} P(o|\mathbf{o}) \sum_{\mathbf{E}' \in \mathcal{E}^*} [p^{\mathbf{o}o}_{\mathbf{E}'} \cdot V_{\mathbf{o}o}(\mathbf{E}')] \\ &= -c(E) + \sum_{o \in \mathcal{O}(E)} P(o|\mathbf{o})\, V_{\mathbf{o}o}(\sigma) \\ &= R(\mathbf{o}, E) + \sum_{o \in \mathcal{O}(E)} T(\mathbf{o}, E, \mathbf{o}o)\, V_{\mathbf{o}o}(\sigma) \\ &= R(s, \sigma(s)) + \sum_{s' \in S} T(s, \sigma(s), s')\, V_{s'}(\sigma). \end{aligned}$$

Either way, (C.1) is satisfied. □

This allows the use of policy iteration to solve an informational decision problem with a finite horizon (Sutton and Barto, 1998). Moreover, the Bellman equation with a max operator over actions, as used in the value iteration algorithm, can be shown to hold for the optimal strategy by a slightly longer argument (Hay and Russell, 2011). Of course, for any large problem, this means computing the optimal policy is intractable. In an important recent paper, Krause and Guestrin (2009) have developed algorithms for efficiently finding optimal policies in restricted settings, specifically when the underlying probabilistic model forms a *chain graphical model*. However, they also show that the problem in general is $\mathbf{NP}^{\mathbf{PP}}$-hard. $\mathbf{NP}^{\mathbf{PP}}$ is the class of problems solvable in non-determinstic polynomial time with an oracle for counting problems solvable in polynomial time. This complexity class is quite typical of AI planning problems, and it means approximate techniques will be necessary for efficient computation in any sufficiently complex domain.

C.2 Omniscience

Save for Good (1983) and Skyrms (1990), every response to Savage's comments I have been able to find in the philosophical literature is directed at the problem of omniscience. That is, they focus on the question of how to characterize the gambler's uncertainty about the remote digit of π, while also respecting the fact that the agent has knowledge sufficient to deduce the identity of that digit.

Beginning with Hacking's original response (Hacking, 1967), a popular line has been to introduce *impossible worlds* in which true mathematical or logical statements can come out false. This response has not been absent progress either. Lipman (1991), directly inspired by Savage's discussion, and using impossible worlds, shows how to construct a single probability space representing all of the agent's uncertainty, including uncertainty about how to reduce its own uncertainty, as the least fixed point of a constructive process iterated into the transfinite. He takes this to be a partial "solution" to the regress problem (Lipman, 1991, 1106), with the goal of capturing what the agent's own "perception" of the deliberative situation looks like. Interesting as this is, in the end the agent is assumed to maximize expected utility in this enormous model, which one might worry leads us right back to the problems discussed above in §4.5.

As an alternative to the impossible worlds approach, in recent work picking up on a different thread from Hacking (1967), Gaifman (2004) argues that mathematical inquiry ought to be treated on a par with empirical inquiry. After all, mathematicians make conjectures, assess their likelihood, decide accordingly on strategies, and so on. Gaifman shows how logical or mathematical reasoning can be combined with probabilistic reasoning, so that, e.g., having argued that some statement holds with high probability—say, that a given number is prime, using a probabilistic test—one can deduce further consequences that are guaranteed to hold with probability at least as high. As Gaifman acknowledges, the justification for using probabilistic methods to investigate decidable mathematical problems "derives from the limited resources, which amounts in this case to considerations of cost. The basic principle, argued for by Hacking, is that costs of computation and costs of empirical inquiry are on a par" (Gaifman, 2004, 113). This was precisely the intuition behind our "internalizing" the value of information framework to deal with computation. However, in Gaifman's paper, like in Hacking's paper, this aspect of bounded mathematical reasoning is left informal and anecdotal.

Even more recently, Seidenfeld et al. (2012) argue against some of the other treatments of omniscience in the literature, including those by Hacking (1967) and Gaifman (2004). They also complain that dealing with Savage's challenge as Good suggested, using dynamic probabilities, is unprincipled and flouts the Principle of Total Evidence. Instead, they favor their own treatment of the puzzle in terms of *degrees of incoherence*, allowing a more fine-grained taxonomy of belief states between coherence and incoherence. They show how an agent can perform numerical calculations resembling statistical inference to learn about mathematical truths, in such a way that the degree of incoherence is guaranteed not to increase, and may well decrease.

While a thorough discussion of these proposals is beyond the scope of the current chapter, it is worth explaining how these issues fit together, and why some of this work ought not be seen as in conflict with the approach taken here. The approach to bounded rationality and reasoning we have pursued in this chapter, and in this dissertation more generally, is initially neutral about the characterization of an agent's mental state. In our general treatment, we have assumed only that the agent's "program" can be associated with a probability distribution over actions (or action utilities, as in this chapter), and with a cost function. More specific questions about a particular agent's level of rationality or fitness will of course require specifying more about the agent's architecture (and environment, etc.). Many familiar computational approaches to bounded rationality in the literature—e.g., Rubinstein (1986) on game-theoretic agents as finite automata, which do not deal with omniscience so much as assume it away—fit naturally into this framework. However, we are interested in *human* agents, who may reason or deliberate about any topic whatsoever, including mathematical subjects, as Savage's example makes salient. This suggests that if we want to apply the general framework to more sophisticated agents, we need to know how they see the world and what their computations consist in, which seems to require facing the omniscience problem head-on.

We can view the proposals by Lipman, Gaifman, Seidenfeld et al., and others, as offering ways of characterizing agents' mental states, including how they see the world, and in some cases (Gaifman, 2004; Seidenfeld et al., 2012) as offering suggestions about how agents so characterized *ought* to reason. I find this an important project, and it may well lead to useful insights into mathematical and other types of formal reasoning. However, even with an adequate way of representing agents' epistemic situations (e.g., with respect to mathematical statements), we still face the problem of understanding how such an agent should reason *in real time*, *under realistic resource constraints*, as well as how the agent should reason about its own resources, if at all. These are precisely the problems we have sought to understand with the value of information account. While these authors assiduously acknowledge the importance of bounds and costs, they play little role in the actual analyses. Consequently, they also do not tell us anything about *when* or *how* to deliberate.

It may be profitable to combine one of these analyses with the approach taken in this chapter. That said, the question of how people see the world, and how we as theorists should characterize how they see the world, I take to be in large part empirical (notwithstanding well known claims by Lewis 1974 and Davidson 1975 to the contrary). The view of subjective probability put forth in Chapter 2 proposes that the way people deal with uncertainty in many typical cases is not by harboring explicit probability representations, but by sampling from implicitly represented distributions, e.g., as given by generative processes. This view does not preclude our characterizing agents using full probability spaces, perhaps as a first approximation. But approximations need to be taken with a grain of salt. Especially in order to understand bounded rationality, and how real agents deal with costs and resource bounds, we must work at this more exact, concrete, *algorithmic* level, as argued in Ch. 3. The Sampling Hypothesis raises distinctive questions about bounded rationality—and together with the methodological framework of (boundedly) rational analysis, suggests answers to these questions—as we have explored in some depth. It sidesteps the omniscience problem altogether, since an agent is only assumed to be capable of drawing (possibly biased) samples and using these samples to make a decision. This hypothesis is attractive in part because it is nonetheless closely tied to the more familiar (static) probabilistic representations of agents' mental states; that is, a sampling agent is typically assumed to be sampling from some distribution. But any such proposal—where we require concrete, implementable *algorithms*, so that agents have to *do* something to reason—will not suffer from the logical omniscience problem.

One might therefore worry that these some of these other purported treatments of omniscience are attempting to solve a problem that is not there.

Whatever the case, we have not said much of anything in this dissertation about explicit reasoning with probabilities, or about explicit mathematical or logical reasoning. In that sense, we cannot claim serious progress on the omniscience problem as it is usually understood. But, as I hope to have made clear, this is a separate issue from Savage's second challenge, our main topic in §4: to show how an agent might reason about its own reasoning, taking costs into account, so as to improve its ability to take smarter actions, given limited time and resources, without falling into regress.

Bibliography

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Lawrence Earlbaum Associates, Inc.

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

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences 22*(4), 577–609.

Baxter, J. (1997). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning 28*, 7–39.

Ben-David, S., B. Chor, O. Goldreich, and M. Luby (1992). On the theory of average case complexity. *Journal of Computer and System Sciences 44*, 193–219.

van Benthem, J. F. A. K. (2007). Computation as conversation. In S. B. Cooper, B. Löwe, and A. Sorbi (Eds.), *New Computational Paradigms: Changing Conceptions of What is Computable*, pp. 35–58. Springer.

Bernardo, J. M. and A. M. Smith (1994). *Bayesian Theory.* John Wiley and Sons.

Boghossian, P. (2014). What is inference? *Philosophical Studies*. forthcoming.

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

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason.* Harvard University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence 47*, 139–159.

Broome, J. (2013). *Rationality through Reasoning.* Wiley Blackwell.

Buesing, L., J. Bill, B. Nessler, and W. Maass (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology 7*(11).

Canini, K. R., M. M. Shashkov, and T. L. Griffiths (2010). Modeling transfer learning in human categorization with the Hierarchical Dirichlet Process. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 151–158.

Chi, M. T. H., M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science 13*, 145–182.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* MIT Press.

Churchland, P. S. and T. J. Sejnowski (1994). *The Computational Brain.* MIT Press.

Clark, A. (1997). *Being There: Putting Mind, Body, and World Together Again.* MIT Press.

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence 169*, 104–141.

Craik, K. (1943). *The Nature of Explanation.* Cambridge University Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater and M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*, pp. 59–75. Oxford University Press.

Danks, D. and F. Eberhardt (2009). Explaining norms and norms explained [commentary]. *Behavioral and Brain Sciences 32*(1), 86–87.

Davidson, D. (1975). Hempel on explaining action. *Erkenntnis 10*(3), 239–253.

Daw, N. D., A. C. Courville, and P. Dayan (2008). Semi-rational models of conditioning: the case of trial order. In N. Chater and M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, pp. 431–452. Oxford University Press.

Daw, N. D., Y. Niv, and P. Dayan (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavior control. *Nature Neuroscience 8*(12), 1704–1711.

Dayan, P. and N. D. Daw (2008). Connections between computational and neurobiological perspectives on decision making. *Cognitive, Affective, & Behavioral Neuroscience 8*(4), 429–453.

Dean, T. and M. Boddy (1988). An analysis of time-dependent planning. In *Proceedings of AAAI*, pp. 49–54.

DeGroot, M. H. (2004). *Optimal Statistical Decisions.* John Wiley & Sons.

Denison, S., E. Bonawitz, A. Gopnik, and T. L. Griffiths (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition 126*, 285–300.

Dennett, D. C. (1981a). *Brainstorms: Philosophical Essays on Mind and Psychology.* MIT Press.

Dennett, D. C. (1981b). Three kinds of intentional psychology. In R. Healey (Ed.), *Reduction, Time, and Reality*, pp. 37–61. Cambridge University Press.

Dewey, J. (1922). *Human Nature and Conduct.* Henry Holt & Co.

Domingos, P. and D. Lowd (2009). *Markov Logic: An Interface Layer for Artificial Intelligence.* Morgan and Claypool.

Douven, I. (2002). Decision theory and the rationality of further deliberation. *Economics and Philosophy 18*(2), 303–328.

Eberhardt, F. and D. Danks (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines 21*(3), 389–410.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo Method. *Los Alamos Science 15*, 131–137.

Eells, E. (2005). Confirmation theory. In S. Sarkar and J. Pfeifer (Eds.), *The Philosophy of Science: An Encyclopedia*, pp. 144–150. Routledge.

Eriksson, L. and A. Hájek (2007). What are degrees of belief? *Studia Logica 86*(2), 183–213.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association 67*(337), 81–102.

de Finetti, B. (1974). *Theory of Probability*, Volume 1. Wiley, New York.

Fiser, J., P. Berkes, G. Orbán, and M. Lengyel (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Science 14*(3), 119–130.

Fodor, J. A. (1983). *The Modularity of Mind.* MIT Press.

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In *Pylyshyn (1987)*, pp. 139–149.

Frank, M. C. and N. D. Goodman (2012). Predicting pragmatic reasoning in language games. *Science 336.*

Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly 15*(2), 157–162.

Freer, C., D. Roy, and J. Tenenbaum (2012). Towards common-sense reasoning via conditional simulation: Legacies of Turing in artificial intelligence. In R. Downey (Ed.), *Turing's Legacy.* ASL Lecture Notes in Logic.

Frege, G. (1979). Logic. In *Posthumous Writings*. University of Chicago Press.

Gaifman, H. (2004). Reasoning with limited resources and assigning probabilities to arithmetical statements. *Synthese 140*, 97–119.

Galton, F. (1889). *Natural Inheritance*. MacMillan.

Gershman, S. J. and N. D. Daw (2012). Perception, action, and utility: the tangled skein. In M. Rabinovich, K. Friston, and P. Varona (Eds.), *Principles of Brain Dynamics: Global State Interactions*, pp. 293–312. MIT Press.

Gershman, S. J., E. Vul, and J. B. Tenenbaum (2012). Multistability and perceptual inference. *Neural Computation 24*, 1–24.

Gigerenzer, G. and D. G. Goldstein (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review 103*(4), 650–699.

Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Good, I. J. (1960). Subjective probability as the measure of a non-measurable set. In E. Nagel, P. Suppes, and A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, pp. 319–329.

Good, I. J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science 17*, 319–321.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press.

Goodman, N. D., V. K. Mansinghka, D. Roy, K. Bonawitz, and J. B. Tenenbaum (2008). Church: A language for generative models. In *Uncertainty in Artificial Intelligence*, Volume 22. AUAI Press.

Goodman, N. D., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths (2008). A rational analysis of rule-based concept learning. *Cognitive Science 32*, 108–154.

Gopnik, A., C. Glymour, D. Sobel, L. Schulz, T. Kushnir, and D. Danks (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review 111*(1), 3–32.

Gould, S. J. and R. C. Lewontin (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society, B 205*(1161), 581–598.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics*, Volume 3. Academic Press.

Griffiths, T. L., F. Lieder, and N. D. Goodman (2014). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*. forthcoming.

Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science 34*(4), 311–325.

Harman, G. (1986). *Change in View.* MIT Press.

Hay, N. J. and S. Russell (2011). Metareasoning for Monte Carlo tree search. Technical Report No. UCB/EECS-2011-119, UC Berkeley.

Hay, N. J., S. Russell, S. E. Shimony, and D. Tolpin (2012). Selecting computations: Theory and applications. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pp. 346–355.

Holton, R. (2014). Intention as a model for belief. In M. Vargas and G. Yaffe (Eds.), *Rational and Social Agency: Essays on the Philosophy of Michael Bratman.* Oxford University Press.

Horvitz, E., J. Breese, and M. Henrion (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning 2*, 247–302.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics (SSC-2)*, 22–26.

Humphreys, P. (1994). Numerical experimentation. In P. Humphreys (Ed.), *Patrick Suppes: Scientific Philosopher*, Volume 2. Kluwer.

Icard, T. F. (2012). Inclusion and exclusion in natural language. *Studia Logica 100*(4), 705–725.

Icard, T. F. (2013). The place of logic in a probabilistic semantics: Comments on Goodman. In J. van Benthem and F. Liu (Eds.), *Logic Across the University: Foundations and Applications*, pp. 351–362. College Publications, Studies in Logic.

Icard, T. F. (2014). Toward boundedly rational analysis. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting in Cognitive Science*, pp. 637–642.

Icard, T. F. and L. S. Moss (2014). Recent progress on monotonicity. *Linguistic Issues in Language Technology*. forthcoming.

Icard, T. F., E. Pacuit, and Y. Shoham (2010). Joint revision of belief and intention. In *Proceedings of Knowledge Representation and Reasoning.*

James, W. (1890). *The Principles of Psychology.* Henry Holt & Co.

James, W. (1907). What pragmatism means. In *Pragmatism, a new name for some old ways of thinking.* Longmans, Green, and Company.

Jeffrey, R. C. (1965). *The Logic of Decision.* McGraw Hill.

Jones, M. and B. C. Love (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences 34*(4), 169–231.

Kadane, J. B., M. Schervish, and T. Seidenfeld (2008). Is ignorance bliss? *Journal of Philosophy 105*(1), 5–36.

Kemp, C., N. D. Goodman, and J. B. Tenenbaum (2010). Learning to learn causal models. *Cognitive Science 34*(7), 1285–1243.

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

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Krause, A. and C. Guestrin (2009). Optimal value of information in graphical models. *Journal of Artificial Intelligence Research 35*, 557–591.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Science 14*, 293–300.

Kwisthout, J., T. Wareham, and I. van Rooij (2008). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science 35*, 779–784.

Lagoudakis, M. G., M. L. Littman, and R. E. Parr (2001). Selecting the right algorithm. In *Proceedings of AAAI*, pp. 74–75.

Lee, M. D. and T. D. R. Cummins (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review 11*(2), 343–352.

Levy, R., F. Reali, and T. L. Griffiths (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems 21*, 937–944.

Lewandowsky, S., T. L. Griffiths, and M. L. Kalish (2009). The wisdom of individuals: Exploring peoples knowledge about everyday events using iterated learning. *Cognitive Science 33*, 969–998.

Lewis, D. K. (1974). Radical interpretation. *Synthese 23*, 331–344.

Lieder, F., T. L. Griffiths, and N. D. Goodman (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems 25*, 2699–2707.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics 27*, 986–1005.

Lipman, B. (1991). How to decide how to decide how to . . . : Modeling limited rationality. *Econometrica 59*, 1105–1125.

Lochmann, T. and S. Deneve (2011). Neural processing as causal inference. *Current Opinion in Neurobiology 21*, 774–781.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* John Wiley & Sons.

Luce, R. D. and P. Suppes (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. H. Galanter (Eds.), *Handbook of Mathematical Psychology*, Volume 3, pp. 249–410. Wiley.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

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

Marcus, G. F. and E. Davis (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*. (forthcoming). DOI: 10.1177/0956797613495418.

Marr, D. (1982). *Vision.* W.H. Freeman and Company.

Marr, D. and T. Poggio (1976). From understanding computation to understanding neural circuitry. MIT A.I. Memo 357.

McCarthy, J. (1959). Programs with common sense. In *Symposium Proceedings on Mechanisation of Thought Processes*, Volume 1, pp. 77–84.

McFadden, D. L. (1973). Conditional logic analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics.* Academic Press.

Millgram, E. (1991). Harman's hardness arguments. *Pacific Philosophical Quarterly 72*(3), 181–202.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories.* MIT Press.

Minsky, M. L. (1965). Matter, mind, and models. In *Proceedings of the International Federation of Information Processing*, Volume 1, pp. 45–49.

Moreno-Bote, R., D. C. Knill, and A. Pouget (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences 108*(30), 12491–12496.

Movellan, J. R. and J. L. McClelland (2001). The Morton-Massaro Law of Information Integration: Implications for models of perception. *Psychological Review 108*, 113–148.

Mozer, M. C., H. Pashler, and H. Homaei (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science 32*, 1133–1147.

Neisser, U. (1967). *Cognitive Psychology.* Prentice Hall.

Nelson, T. O. and L. Narens (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation 26*, 125–173.

Newell, A. and H. A. Simon (1976). Computer science as empirical inquiry: Symbols and search. *Communication of the ACM 19*(3), 113–126.

Nozick, R. (1981). *Philosophical Explanations.* Harvard University Press.

Oaksford, M. and N. Chater (2007). *Bayesian Rationality.* Oxford University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.

Perfors, A. (2012). Bayesian models of cognition: What's built in after all? *Philosophy Compass 7*(2), 127–138.

Pierce, C. S. (1887). Logical machines. *American Journal of Psychology 1*, 165–170.

Pylyshyn, Z. W. (1984). *Computation and Cognition.* MIT Press.

Pylyshyn, Z. W. (Ed.) (1987). *The Robot's Dilemma.* Ablex.

Raiffa, H. and R. Schlaifer (1961). *Applied Statistical Decision Theory.* Harvard Studies in Managerial Economics.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *Foundations of Mathematics and Other Logical Essays.* Martino Fine.

Reichenbach, H. (1949). *The Theory of Probability.* University of California Press.

van Rooij, R. (2004). Utility, informativity, and protocols. *Journal of Philosophical Logic 33*(4), 389–419.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory 39*, 83–96.

Rumelhart, D. E., J. L. McClelland, and The PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press.

Russell, S. and D. Subramanian (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research 2*, 1–36.

Russell, S. and E. Wefald (1991). *Do the Right Thing: Studies in Limited Rationality.* MIT Press.

Ryle, G. (1949). *The Concept of Mind.* University of Chicago Press.

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

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

Sanborn, A. N., V. K. Mansinghka, and T. L. Griffiths (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review 120*, 411–437.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons.

Savage, L. J. (1967). Difficulties in the theory of personal probability. *Philosophy of Science 34*(4), 305–310.

Schacter, D. L., D. R. Addis, and R. L. Buckner (2008). Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences 1124*, 39–60.

Schwarz, N., H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka, and A. Simons (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology 61*(2), 195–202.

Seidenfeld, T., M. J. Schervish, and J. B. Kadane (2012). What kind of uncertainty is that? Using personal probability for expressing one's uncertainty about logical and mathematical propositions. *Journal of Philosophy 109*(8/9), 516–533.

Seth, A. K. (1999). Evolving behavioural choice: An exploration of Hernnstein's Matching Law. In D. Floreano, J.-D. Nicoud, and F. Mondada (Eds.), *Proceedings of the Fifth European Conference on Artificial Life*, pp. 225–236. Springer.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics 69*(1), 99–118.

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

Simon, H. A. (1976). From substantive to procedural rationality. In T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, and G. R. Wagenaar (Eds.), *25 Years of Economic Theory*, pp. 65–86. Springer.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (Ed.), *Architectures for Intelligence: The 22nd Carnegie Mellon Symposium on Cognition*, pp. 25–39. Lawrence Earlbaum Associates, Inc.

Simon, H. A. (1995). Artificial intelligence: an empirical science. *Artificial Intelligence 77*, 95–127.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Solway, A. and M. M. Botvinick (2012). Goal-directed decision making as probabilistic inference. *Psychological Review 119*(1), 120–154.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist 57*(4), 421–457.

Stewart, N., N. Chater, and G. D. Brown (2006). Decision by sampling. *Cognitive Psychology 53*, 1–26.

Stich, S. P. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Cognitive Evaluation*. MIT Press.

Stratonovich, R. L. (1965). On value of information. *Izvestiya USSR Academy of Sciences, Technical Cybernetics 5*, 3–12.

Sundareswara, R. and P. Schrater (2007). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision 8*(5), 1–19.

Suppes, P. (1974). The measurement of belief. *The Journal of the Royal Statistical Society, Series B 36*(2), 160–191.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning*. MIT Press.

Tenenbaum, J. T., C. Kemp, T. L. Griffiths, and N. D. Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science 331*, 1279–1285.

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun and L. Pratt (Eds.), *Learning to Learn*, pp. 181–209. Kluwer.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*(4), 273–286.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*, 433–460.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.

Tversky, A. and D. Kahneman (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review 90*(4), 293–315.

Vilares, I. and K. P. Kording (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences 1224*, 22–39.

Vul, E. (2010). *Sampling in Human Cognition.* Ph. D. thesis, MIT.

Vul, E., N. D. Goodman, T. L. Griffiths, and J. B. Tenenbaum (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pp. 148–153.

Vul, E., N. D. Goodman, T. L. Griffiths, and J. B. Tenenbaum (2013). One and done? Optimal decisions from very few samples. *Cognitive Science*. forthcoming.

Vul, E. and H. Pashler (2008). Measuring the crowd within. *Psychological Science 19*(7), 645–647.

Weirich, P. (2004). *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances.* Oxford University Press.

Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine.* MIT Press.

Yang, T. and M. N. Shadlen (2007). Probabilistic reasoning by neurons. *Nature 447*, 1075–1082.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science 67*(1), 45–69.