BAYESIAN EPISTEMOLOGY AND HAVING EVIDENCE

A Dissertation Presented

by

JEFFREY STEWART DUNN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2010

Department of Philosophy

BAYESIAN EPISTEMOLOGY AND HAVING EVIDENCE

A Dissertation Presented

by

JEFFREY STEWART DUNN

Approved as to style and content by:

_____
Hilary Kornblith, Chairperson

_____
Phillip Bricker, Member

_____
David Christensen, Member

_____
Christopher Meacham, Member

_____
Shlomo Zilberstein, Member

_____
Phillip Bricker, Head
Department of Philosophy

DEDICATION

To Helen.

ACKNOWLEDGMENTS

In writing this dissertation I have been helped in many ways and by many people. To each of these people I am extremely grateful.

First and foremost, I express my deepest gratitude to my advisor, Hilary Kornblith. Hilary is everything that an advisor should be. He always made himself available for discussion, he provided me with endless good advice, he challenged me philosophically, and he showed me how many things outside philosophy are relevant to philosophy. On top of all that he possesses an uncanny ability to return drafts—with comments!—in record time. I thank him for all he has done.

I also sincerely thank Chris Meacham. Chris was always ready to discuss my new ideas, and provide invaluable feedback. He patiently taught me about how to use formal methods, and corrected my many errors and false starts as I was getting the hang of it. I also thank David Christensen and Phil Bricker for challenging and helpful comments on many parts of this dissertation. I thank Shlomo Zilberstein for serving on my committee and for introducing me to computer science.

I am also grateful to those faculty members who weren't on my committee, but provided help on the dissertation or who were especially instrumental in helping me learn how to do philosophy. I thank Fred Feldman, Jonathan Schaffer, Kevin Klement, Casey Perin, Ed Gettier, Pete Graham, Brad Skow, Lynne Baker, Joe Levine, and Louise Antony.

I am especially grateful to Beth Grybko for her tireless help with the administrative aspects of successfully navigating graduate school and completing a dissertation.

In writing this disseration I have also benefited from many discussions with fellow graduate students at UMass. I especially thank Kristoffer Ahlstrom, Indrani Bhattacharjee, Jim Binkoski, Heidi Buetow, Donovan Cox, Sam Cowling, Jeremy Cushing, Lowell Friesen, Jayme Johnson, Namjoong Kim, Justin Klocksiem, Casey Knight, Barak Krakauer, Uri Leibowitz, Peter Marchetto, Meghan Masto, Kirk Michaelian, Josh Moulton, Kristian Olsen, James Platt, Alex Sarch, and Kelly Trogdon.

Ed Ferrier and Einar Bohn deserve special thanks. They provided much helpful philosophical discussion, but it is their friendship, sense of humor, and their love of Fitzwilly's that made my time at graduate school much better than it would have otherwise been.

I am also deeply grateful to my parents for their love and their enthusiastic support of my decision to go to graduate school. The environment of learning that they created in our home has had a significant impact on who I am, my interest in philosophy, and on my ability to complete this dissertation. I also thank my parents-in-law for their generosity, support, and kindness.

Finally, my deepest gratitude is owed to my wife, Helen. I would not have completed this dissertation were it not for her kind and loving support. She has picked me up when things looked grim and celebrated with me when things were going well. I look forward to enjoying the future ups and downs of life with her, and I sincerely thank her for all her love and support.

ABSTRACT

BAYESIAN EPISTEMOLOGY AND HAVING EVIDENCE

SEPTEMBER 2010

JEFFREY STEWART DUNN, B.A., WASHINGTON STATE UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hilary Kornblith

Bayesian Epistemology is a general framework for thinking about agents who have beliefs that come in degrees. Theories in this framework give accounts of rational belief and rational belief change, which share two key features: (i) rational belief states are represented with probability functions, and (ii) rational belief change results from the acquisition of evidence. This dissertation focuses specifically on the second feature. I pose the *Evidence Question*: What is it to have evidence?

Before addressing this question we must have an understanding of Bayesian Epistemology. The first chapter argues that we should understand Bayesian Epistemology as giving us theories that are evaluative and not action-guiding. I reach this verdict after considering the popular 'ought'-implies-'can' objection to Bayesian Epistemology.

The second chapter argues that it is important for theories in Bayesian Epistemology to answer the Evidence Question, and distinguishes between internalist and externalist answers.

The third and fourth chapters present and defend a specific answer to the Evidence Question. The account is inspired by reliabilist accounts of justification, and

attempts to understand what it is to have evidence by appealing solely to considerations of reliability. Chapter 3 explains how to understand reliability, and how the account fits with Bayesian Epistemology, in particular, the requirement that an agent's evidence receive probability 1. Chapter 4 responds to objections, which maintain that the account gives the wrong verdict in a variety of situations including skeptical scenarios, lottery cases, scientific cases, and cases involving inference. After slight modifications, I argue that my account has the resources to answer the objections.

The fifth chapter considers the possibility of losing evidence. I show how my account can model these cases. To do so, however, we require a modification to Conditionalization, the orthodox principle governing belief change. I present such a modification.

The sixth and seventh chapters propose a new understanding of Dutch Book Arguments, historically important arguments for Bayesian principles. The proposal shows that the Dutch Book Arguments for implausible principles are defective, while the ones for plausible principles are not.

The final chapter is a conclusion.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

INTRODUCTION

What is evidence? This seems like an important question. Evidence is that on which we base our convictions—or acquittals—of those accused of crimes in criminal trials. Evidence plays a similarly important role in civil lawsuits. Currently, there is a push for so-called "evidence-based medicine". Climate scientists have recently come under criticism for being less than forthcoming about the evidence in favor of global warming.

Evidence is clearly important to our practical pursuits. This lends interest to questions about the nature of evidence. But evidence is also important in more theoretical pursuits, specifically in epistemology. According to some views in epistemology, evidence plays an indispensable role.[1] Other views do not give evidence such a central place.[2] But given its prominence in epistemological thinking, even these views must say something about what evidence is.

The overarching goal of this dissertation is to say something about what evidence *is*. Specifically, I seek to answer the *Evidence Question*: What is it to have evidence? Stated in such a general way, this question is too broad to be intelligibly answered. We need some way to restrict the question to a manageable size. To do this, I will adopt the methodology of answering the question from a particular theoretical perspective: the perspective of Bayesian Epistemology.[3] There are three main reasons for choosing this perspective. First, Bayesian Epistemology makes use of evidence as a central notion. This

1 For instance, Evidentialism (Feldman & Conee [1985]) and Bayesian Epistemology (see citations in footnote 3).
2 For instance, standard Process Reliabilism, of the sort defended in Goldman ([1979]).
3 Representative works in this area include Ramsey ([1926/1990]), Jeffrey ([1965]), Howson & Urbach ([1989], [1993]), Kaplan ([1998]), Howson ([2003]), Bovens & Hartmann ([2003]), and Christensen ([2004]).

makes it a well-suited perspective from which to investigate the nature of evidence. Second, Bayesian Epistemology is a relatively well-worked out and successful approach to epistemological theorizing. It offers us the structure of mathematical formalism, which allows us to state things in a clear and precise way. In addition, it gives a broad enough framework so that we can address traditional epistemological problems (such as skepticism), but also more modern epistemological problems (such as those having to do with inductive inference). Third, and finally, Bayesian Epistemology is the dominant theory that countenances *degrees of belief*, in contrast to epistemological theories where *all-or-nothing beliefs* take center stage.[4] I won't do anything to argue that the more fine-grained approach is appropriate, but I think that it is, and this privileges Bayesian Epistemology in my estimation.

So, I will attempt to say something about what evidence is and to answer The Evidence Question, but will do this from the theoretical perspective of Bayesian Epistemology. Some may see the prominence of this theoretical perspective as a disadvantage. But I view a theoretical perspective as necessary to such an investigation. In Chapter 1 I will sketch out the basic features of this theoretical perspective. I will also argue in favor of a certain way of understanding the theories that we get from Bayesian Epistemology. This understanding is critical to the rest of the dissertation. In Chapter 2, I begin to investigate what evidence *is*, by asking whether or not the evidence an agent *has* is determined solely by things internal to that agent, or whether external facts and features can determine the evidence an agent has. I argue in favor of allowing some

4 There are many, many works in this tradition, but fairly recent representative works in epistemology include Goldman ([1986]), Conee & Feldman ([2004]), and BonJour ([1985]). There is also work outside of mainstream epistemology exhibiting this general representation of belief states, including work on knowledge bases and work on epistemic logics.

external facts to determine the evidence an agent has, and argue that this *externalist* perspective is consonant with the standard principles of Bayesian Epistemology. In Chapters 3 and 4 I present a novel account of what it is for an agent to have evidence, which I call the Reliabilist Account of Evidence (RAE). The basic idea is that the evidence an agent has is determined by the highly reliable processes of belief formation that are available to the agent. In Chapter 3 I present and clarify the view being proposed, and in Chapter 4 I respond to objections. In Chapters 5-7, I show how, with an account of evidence like RAE, one can address some interesting issues in epistemology. In Chapter 5 I show how we can model agents that *lose* evidence, something that has been tricky to model from a Bayesian perspective. In Chapter 6 I address Dutch Book Arguments, and argue that a clear understanding of what it is to have evidence can help us sort the good Dutch Book Arguments from the bad ones. Finally, in Chapter 7, I discuss what it is to have evidence about one's own doxastic state.

Table 1: Guide to Notation

Notation	**Explanation**
$\wedge$	Truth-functional conjunction.
$\vee$	Truth-functional disjunction.
$\rightarrow$	Truth-functional material implication.
$\neg$	Truth-functional negation.
Uppercase italicized letters	These represent propositions.
$\mathrm{cr}(\bullet)$	A credence (or: degree of belief) function.
$\mathrm{cr}(\bullet \mid P)$	A conditional credence function, in this case, conditional on the proposition P.
$\langle \mathrm{cr}(P) = n \rangle$	The proposition that the credence in proposition P is n. It is useful to use these angled brackets when a proposition about a credence value is the object of a credence, e.g., $\mathrm{cr}(\langle \mathrm{cr}(P) = n \rangle) = n$.

CHAPTER 1

BAYESIAN EPISTEMOLOGY AND 'OUGHT'-IMPLIES-'CAN'

1.1 Introduction

Bayesianism presents us with a general framework for thinking about agents with partial beliefs. One of the most persistent criticisms of this general framework centers on the level of idealization that seems to be employed. Such worries are quite widespread.

In this chapter, I will be addressing a particular form that this worry takes: that Bayesianism gives us theories that violate some epistemic version of an 'ought'-implies-'can' principle. Alvin Goldman ([1978]) tells us, "As in the ethical sphere, 'ought' implies 'can'. Traditional epistemology has often ignored this precept." (p. 510). Goldman is not alone in this sentiment. Many have thought that some form of an 'ought'-implies-'can' (OIC) principle should guide inquiry in epistemology. The worry is that Bayesianism egregiously violates such a principle. Bayesianism provides us with theories of rationality and theories of rationality give us obligations. The obligations given by a Bayesian theory, however, are impossible to meet. So, there is a problem with Bayesianism. In the end, I will conclude that this line of argument fails, but not before revealing something interesting about how to understand the Bayesianism framework.

1.2 Bayesian Epistemology and Rationality

Bayesianism gives us a framework. Specific accounts in the Bayesian family often differ in the details, but they are in agreement about several fundamental issues.

Shared is the idea that an adequate epistemology must be one that recognizes the possibility of *degrees* of belief. Thus, an agent's *epistemic* state, according to Bayesianism is represented by a function that assigns real numbers to propositions (henceforth: "credence function"). Further, an agent's *evaluative* state is represented by a function that assigns real numbers to possibilities (henceforth: "utility function"). Intuitively, the higher the number assigned to a proposition by the credence function, the stronger the belief in that proposition, and the higher the number assigned to a possibility by the utility function, the more that possibility is desired. This is how agents are represented by standard Bayesianism theories.[1]

Standard Bayesian theories then make at least the following three claims about *rational* agents:

PROB: A rational agent's credence function is a probability function. A credence function is a probability function just in case: (i) for all P, $0 \leq \text{cr}(P) \leq 1$, (ii) for all logical truths $\top$, $\text{cr}(\top) = 1$, and (iii) $\text{cr}(P \vee Q) = \text{cr}(P) + \text{cr}(Q)$ for all P, Q such that $(P \wedge Q)$ is contradictory.[2]

COND: A rational agent updates her credences according to conditionalization. That is, $\text{cr}_{new}(\bullet) = \text{cr}_{old}(\bullet|E)$, where '$\text{cr}_{old}(\bullet)$' is the agent's old credence function

[1] There is some variation, however. For example, some represent an agent's epistemic state with a *set* of functions, rather than one function. See, for instance, Kaplan ([1996], [2009]), and Sturgeon ([*forthcoming*]).

[2] These three conditions are essentially the Kolmogorov probability axioms. The third condition is finite additivity. Some opt for countable additivity, where additivity holds for disjunctions of (countably) infinite disjuncts. Since this introduces complications unrelated to my project, I ignore this issue here.

before learning evidence, E, '$cr_{new}(\bullet)$' is the agent's new credence function after learning all and only E, and $cr(A|E) = cr(A \wedge E)/cr(E)$ when $cr(E) > 0$.[3]

EUMAX: a rational agent acts so as to maximize subjective expected utility.[4]

These three conditions are taken to be at least necessary conditions of rationality.

It is important to note that the theories that we will be concerned with strive to be *normative* theories of rationality. This is to be contrasted with *descriptive* theories. A descriptive theory purports to describe and predict the way in which reasoning agents actually change their beliefs and act. A normative theory, on the other hand, purports to tell us something about how agents *should* rationally change their beliefs and act.

There is one more distinction to be drawn before we consider the OIC argument. When talking about rationality there is a plausible and well-known distinction to be drawn between *practical rationality* and *epistemic rationality*. As a rough attempt to make this distinction clear, practical rationality has to do with the choices an agent makes, and epistemic rationality has to do with the agent's beliefs. In this dissertation I will focus on epistemic rationality since I am particularly interested in Bayesian *Epistemology*. With respect to epistemic rationality PROB and *COND* are the most relevant. PROB tells us something about a rational agent's belief state at a time, and

---

[3] This is classical conditionalization. There are alternatives. Some, for instance, opt for Jeffrey conditionalization, which does not imply that agents fully believe new evidence. Nevertheless, for the purposes of this paper, such differences will not be important.

[4] Just as is the case with *COND*, there are different ways of specifying how EUMAX is to go. The most popular alternatives are (i) to prescribe maximization of evidential expected utility and (ii) to prescribe maximization of causal expected utility. If Ω is the set of all possibilities, and $U(w)$ is the utility value assigned to possibility $w \in \Omega$, the evidential expected utility of an act, a, is:

$$eEU(a) = \sum_{w \in \Omega} cr(w|a) \cdot U(w)$$

The causal expected utility of an act, a, is:

$$cEU(a) = \sum_{w \in \Omega} CR(a > w) \cdot U(w)$$

COND tells us something about a rational agent's belief state across time. EUMAX, on the other hand, is more obviously relevant to what choices a rational agent makes and so is concerned with practical rationality. Thus, I will henceforth use the term 'Bayesian Epistemology' (abbreviated 'BE') to refer to the set of theories that take *PROB* and *COND* to be necessary conditions in a normative theory of epistemic rationality.

1.3 Survey of Arguments

There are numerous OIC arguments against BE in the literature. What follows is a brief survey.

Richard Samuels, Stephen Stich, and Luc Faucher ([2004]) give and endorse an OIC argument against what they call the 'Standard Picture', which is similar to BE:

> If ought implies can, we are not obliged to reason in ways that we cannot. But the Standard Picture appears to require us to perform reasoning tasks that are far beyond our abilities. …arguments have been developed against the claim, often associated with the Standard Picture, that we ought to revise our beliefs in such a way as to ensure *probabilistic coherence*. Once more, complexity considerations strongly suggest that we cannot satisfy this standard (Osherson, 1996). And if we cannot satisfy the norms of the Standard Picture, then given that ought implies can, it follows that the Standard Picture is not the correct account of the norms of rationality. (Samuels, et. al. [2004], p. 169)

This argument is clearly directed against *PROB*, and so a Bayesian account of epistemic rationality.

Gilbert Harman ([1988]) also gives an OIC argument. Harman takes himself to be arguing against the thesis he calls *Reasoning is conditionalization*: "The updating of probabilities via conditionalization or generalized conditionalization is (or ought to be)

(where "a > *w*" is the subjunctive conditional "if it were that a, then it would be that *w*"). However, as before, such subtleties will not be important for the purposes of this paper.

the only principle of reasoned revision." ([1988], p. 25) Thus, his argument is directed against *COND*, and so a Bayesian account of epistemic rationality. He writes:

> One can use conditionalization to get a new probability for *P* only if one has already assigned a prior probability not only to *E* but to *P & E*. 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 their denials. Unhappily, this leads to combinatorial explosion, since the number of such conjunctions is an exponential function of the number of possibly relevant evidence propositions. (Harman [1988], pp. 25-6)

Although Harman never explicitly appeals to an OIC principle, he takes the impossibility of following the procedures of conditionalization to be sufficient to refute *Reasoning is conditionalization*.

Hilary Kornblith ([1992]) also gives an OIC argument:

> Probabilistic rules of inference are computationally intractable. It would thus be absurd to complain that human inference fails to measure up to such a standard of proper reasoning. An acceptable account of how we ought to reason must surely be one which is not as deeply impossible to implement as the problem of combinatorial explosion shows statistical inference to be. A reasonable ideal must, at a minimum, be computationally feasible… (Kornblith [1992], p. 910)

Kornblith here talks about "rules of inference." Thus, he is concerned with epistemic rationality. Further, he seems to presuppose an OIC principle when he says that an acceptable 'ought' must not be impossible to implement.

Finally, Edward Stein ([1998]) gives an OIC argument against BE:

> …according to the standard principle of rationality, the consistency preservation principle is a normative principle of reasoning, but as I have sketched above, humans cannot in fact follow this principle because of the finitary predicament; because it is a mistake to say that one ought to do something that one cannot do, the standard picture of rationality must be wrong. (Stein [1998], p. 248)

Although this argument is directed against consistency preservation, Stein makes it clear that a similar argument is applicable to BE where beliefs come in degrees and are

governed by *PROB*. Again, we have an argument against a Bayesian view of epistemic rationality.[5]

It is clear from this survey of arguments that *if* the target is a Bayesian account of rationality, it concerns its account of *epistemic* rationality. This justifies the focus on *PROB* and *COND*. However, one might think that this is too quick: the target of these arguments is not really the Bayesian account of rationality. Perhaps those that give OIC arguments against BE are only arguing that BE doesn't give *practical advice* to agents about how to reason. In this case, it would be wrong to interpret the OIC argument as attacking *PROB* and *COND* as necessary conditions of rationality. Instead, it would be attacking the view that these constraints give practical advice.

Surely, some OIC arguments *do* seem to have this limited aim. For instance, Isaac Levi says that OIC considerations are "relevant to appraising the *applicability* of a theory of rational probability judgment," ([1970], p. 138) and Goldman, in the quote that opens this paper, seems to say that OIC considerations are relevant when we are giving epistemic *advice*.

Despite this, most OIC arguments, including the ones surveyed here, *are* directed at BE as a theory of rationality, not just its advice-giving potential. Stein ([1998]) says that the argument shows that *the standard view is wrong*. Kornblith ([1992]) says that the argument shows that probabilistic rules of inference are *unacceptable as an account of how we ought to reason*. Harman ([1988]) takes his argument to show that conditionalization is *not reasonable revision of belief*. Samuels,

[5] Other OIC arguments against BE can be found in Goldman ([1978], p. 510), Kitcher ([1992], pp. 83-4), Levi ([1970], p. 138), Cherniak ([1986], p. 113), Hooker ([1994], p. 204), and Chiappe & Vervaeke ([1997], p. 809). Christensen ([2004], p. 157) formulates a clear version of the argument, though does not endorse it.

et. al. ([2004]) use the OIC argument to show that Bayesian accounts are *not the correct account of rationality*. C. A. Hooker interprets Cherniak's OIC argument as "defeating the "raison d'etre" of rationality," (Hooker ([1994], p. 204) and Philip Kitcher tells us that the OIC considerations can lead us to "revise normative principles" (Kitcher [1992], p. 84). Thus, I think it is safe to construe the OIC argument as taking aim at BE's account of epistemic rationality, not simply its ability to furnish advice.

1.4 The Argument

I will now formulate a precise version of the argument. Since we are restricting our attention to *epistemic* rationality, here is a first pass at what such an argument would look like:

(1) If *PROB* and *COND* are necessary conditions of epistemic rationality, then humans ought to conform with *PROB* and *COND*.

(2) If humans ought to conform with *PROB* and *COND* then humans can conform with *PROB* and *COND*.

(3) It is not the case that humans can conform with *PROB* and *COND*.

(C) Thus, it is not the case that *PROB* and *COND* are necessary conditions of epistemic rationality.

This is a standard-looking OIC argument, and yet there is a puzzle: line (3) looks to be clearly false. There is a clear sense in which humans *can* conform with *PROB* and *COND*. If, for instance, one's beliefs happened to evolve in the way that *PROB* and *COND* dictate, then one would in fact conform with *PROB* and *COND*. This is possible (though unlikely), and so line (3) is false. Presumably, this is not what those giving the OIC argument have meant to be defending, and so this representation of the argument must be incorrect.

It is plausible that the OIC principle intended by the defenders of the OIC argument is similar to the OIC principle that Frances Howard-Snyder appeals to in her ([1997]). In that paper, Howard-Snyder defends an ethical OIC principle that applies to actions. She argues that the 'can' in the relevant OIC principle is to be understood in terms of what actions are available to us. She notes that, in the sense that is relevant, she cannot beat Karpov in a game of chess. This is true despite the fact that it is possible for her to move the pieces on the board in the way needed to beat Karpov. Thus, 'can' in the OIC principle requires practical availability.

I think it is plausible that this is the kind of OIC principle that is meant to feature in the OIC arguments against BE. It is not practically available to humans to conform to *PROB* and *COND*. It might happen, but there's no non-lucky way for this to happen. That is, we do not have the capacity to reliably conform to *PROB* and *COND*. We can make this thought clearer by giving the following OIC principle:

> (2′) If humans ought to conform with *PROB* and *COND* then humans have the capacity to reliably bring about conformance with *PROB* and *COND*.

This then gives us the following alternative for line (3):

> (3′) It is not the case that humans have the capacity to reliably bring about conformance with *PROB* and *COND*.

This clarification of the argument does not suffer from simple refutation in virtue of lucky conformance to the principles. In addition, this fits better with the sentiments of those who have offered OIC arguments, claiming that agents do not have the *capacity* to conform with the principles. For instance, this sentiment is present in Harman's ([1988]) presentation of the argument. Read uncharitably, Harman's OIC argument against *COND* is unconvincing. It is true that one needs to have conjunctions

of all one's beliefs if one is to be *ready* to conditionalize. But conformance with *COND* doesn't require that one be *ready* to conditionalize, rather it requires that one *actually* conditionalize. For this to be the case, then, we only need to require that the agent has conjunctions of all his beliefs and all his evidence propositions, not all *possible* evidence propositions. The right response to this kind of challenge is that the constraints of an epistemic theory don't just give one obligations to actually conditionalize, but rather give obligations to be the sort of thing that could conditionalize, independent of the kinds of evidence one may get. That is, if our epistemic theory says that we should conditionalize, then we need to have the capacity to reliably bring this about. Altering the argument according to (2′) and (3′) is thus faithful to the arguments surveyed.

So, the argument we are considering is:

(1) If *PROB* and *COND* are necessary conditions of epistemic rationality, then humans ought to conform with *PROB* and *COND*.

(2′) If humans ought to conform with *PROB* and *COND* then humans have the capacity to reliably bring about conformance with *PROB* and *COND*.

(3′) It is not the case that humans have the capacity to reliably bring about conformance with *PROB* and *COND*.[6]

[6] This premise is supported by considerations about the computational intractability of algorithms guaranteeing conformance with *PROB* and *COND*. It should be pointed out that even if there is no computationally possible *algorithmic* procedure that brings about conformance with *PROB* and *COND*, there still could be a non-algorithmic procedure that does. So, there is a tacit assumption that if there is no implementable algorithmic procedure, then there is no such procedure, full stop. The general idea behind this is that any *non*-algorithmic procedure that guaranteed conformance with *PROB* or *COND* amounts to some mysterious or magical faculty, and it is safe to put such a possibility aside. It is a very interesting question whether or not this assumption is justified, but I do not wish to address that question here. Thus, I will grant the assumption and not discuss it further.

A good presentation of the computational intractability of *PROB* is in Cherniak ([1986]). Cherniak notes that the satisfiability problem—testing a set of sentences for truth-functional consistency—is a member of the class of NP-complete problems. Such problems are generally regarded to be the most difficult class of computational problems. Writes Cherniak: "…perfect capacity even just to make all tautological inferences is the case par excellence of a problem-solving capacity that is strongly conjectured to require computationally intractable algorithms." ([1986], p. 80) If this is true, then there is no algorithmic procedure, implementable in a human-sized brain, guaranteed to bring about truth-functional consistency. But since a set of truth-functionally consistent beliefs is a special case of a probabilistically coherent set of beliefs (the case where beliefs receive only the extreme credence-values

(C) Thus, it is not the case that *PROB* and *COND* are necessary conditions of epistemic rationality.

It is important to note the form that the OIC principle takes in this argument:

(OIC): If humans ought to conform with P, then humans have the capacity to reliably bring about conformance with P.

(OIC) is at least *prima facie* plausible, claiming that conformance with a constraint is not obligatory unless there is a reliable way of bringing about conformance with that constraint. Indeed, there are cases that tell in favor of the principle. Imagine that 50 fair, six-sided dice are about to be tossed in succession in front of Adam. Adam's friend claims that Adam has an obligation to conform to the following principle:

P: Your predictions of the outcomes of the dice tosses are accurate.

It seems plausible that Adam does not have a genuine obligation to conform to P, precisely because it conflicts with some kind of OIC principle. But, of course, there is a sense in which it is *possible* that Adam pick the correct outcome of all the dice, and so conformance with P does not conflict with the kind of OIC principle that appears in (2). On the other hand, conformance with P *does* conflict with the OIC principle behind (2′) and codified in (OIC). According to such a principle, if Adam is obligated to make his predictions accurate, then Adam must have the capacity to reliably bring about accurate predictions. But this is not possible, and so conformance with P conflicts with (OIC).

---

of 0 and 1), this suffices to establish the claim that there is no algorithmic procedure, implementable on a human-sized brain, guaranteed to bring about probabilistic coherence.

If *COND* is the focus, (3′) is justified by showing that anything that is able to calculate new credences in conformance with *COND* and be *ready* to conditionalize on arbitrary evidence propositions must have much more computing power than humans do. For the details of this argument, see Harman ([1988], pp. 25-6).

1.5 The Anti-Voluntarist Problem

Now that the OIC argument has been presented, I will offer two responses. The first response points out that commitment to the OIC principle in line (2′) commits one to other, more objectionable epistemological principles. The second response takes issue with the way in which BE must be understood for the OIC argument to work. I begin with the first of these responses.

To see the difficulty, consider the following argument, which we might call "The Anti-Voluntarist Argument"[7]:

(V1) If we have epistemic obligations, then doxastic attitudes must sometimes be under our voluntary control.

(V2) Doxastic attitudes are never under our voluntary control.

(V3) Thus, we do not have any epistemic obligations.

No defender of the OIC argument we have surveyed should accept the conclusion of this argument. The first premise of the OIC argument claims that necessary conditions of epistemic rationality entail epistemic obligations. (V3) claims that there are no epistemic obligations. Thus, one can defend the OIC argument and accept (V3) only if one thinks that there are *no* necessary conditions of epistemic rationality. Certainly none of the defenders of the OIC argument surveyed here endorse this pessimistic view. And it would be strange for someone who *did* to offer the OIC argument. Premise (1) of the OIC argument together with (V3) make a much neater argument against BE. So, defenders of the OIC argument will not accept (V3). Since the argument is valid, the defender must reject (V1) or (V2). But which premise should be rejected? I will argue that neither option is attractive for the defender of the OIC argument.

1.5.1 Premise (V1)

Consider first (V1). This appears to be supported by a more general principle, which I will call OIVC ("ought implies voluntary control"):

(OIVC): If Xs have an epistemic obligation to believe in some way, then Xs have voluntary control over believing that way.

This principle entails (V1).[8] Given this, rejecting (V1) requires the rejection of (OIVC). I claim that the defender of the OIC argument will have difficulty rejecting (OIVC).

Recall that the relevant principle in the OIC argument is:

(OIC): If humans ought to conform with P, then humans have the capacity to reliably bring about conformance with P.

First note that (OIC) does not *entail* (OIVC). To see this, assume that some agent has the epistemic obligation to believe in a certain way, say, in conformance with principle P. It might be true that it is *possible* that the agent reliably bring about P-conformance (say, there is some practically available and implementable algorithm that will bring about P-conformance), and yet that something (say, implementing the algorithm) is not under voluntary control. However, the converse is true: if I have voluntary control over believing a certain way (say in conformance with principle P), then it is true that I have the capacity to reliably bring about P-conformance. Thus, if (OIVC) is true, then so is

---

[7] A version of this argument can be found in Alston ([1989], p. 118), Plantinga ([1993], p. 38), Steup ([2000]), Feldman ([2001]), Ryan ([2003]), and Chrisman ([2008]).

[8] Assume that (OIVC) is true, and assume that we have epistemic obligations. Plausibly, epistemic obligations are obligations to believe some particular proposition or to conform to some principle of belief-formation/maintenance. But conforming to some principle of belief-formation/maintenance just is to form/maintain your beliefs in ways that conform to the principle. Thus, we can say that all epistemic obligations are obligations to believe *in some way* (where 'some way' may refer to a single belief, or to a set of beliefs – say a *coherent* set, or to a series of belief sets over time). Call this way P. According to (OIVC) (and the assumption that we have epistemic obligations), we have voluntary control over believing in way P. If we have voluntary control over believing in way P, then doxastic attitudes are sometimes under voluntary control (namely, this doxastic attitude, at this time). Thus, (OIVC) entails (V1).

(OIC).[9] Nevertheless, it is *consistent* to embrace (OIC) and deny (OIVC). This shows that if there is a problem with embracing (OIC) and rejecting (OIVC), it is not a problem of inconsistency.

However, even though not inconsistent, it is implausible. There are three reasons for this. First, consider three traditional arguments for OIC principles.[10] The first of these focuses on the action-guiding nature of obligations, noting that we should advise people to fulfill their obligations. However, it makes no sense to advise someone to do what they *cannot* reliably bring about. Note, however, that it equally makes no sense to advise someone to do what is *involuntary*. The second argument focuses on the relation between obligation and blame. If someone violates their obligation, then they are to blame, but it makes no sense to blame someone for failing to do what they *cannot* reliably bring about. However, it equally makes no sense to blame someone for failing to do what is *not within his/her voluntary control*. The third argument starts with the observation that when we learn that someone *cannot* fulfill their obligation, we ask what that person ought to do instead. But note that when we learn that someone has *no voluntary control* over fulfilling their obligation, we similarly ask what they ought to do instead.

I make no claim about the soundness of these arguments. What I do claim is that these are standard arguments in support of OIC principles and that *if* the arguments are

[9] Incidentally, the contrapositive of this gets us Richard Feldman's ([2001]) response to the Anti-Voluntarist Argument: He aims to show that (OIC) is false. Then, since (OIVC) entails (OIC), and since (OIVC) supports (V1), (V1) loses its support. Obviously *this* response to the Anti-Voluntarist Argument is not available to the defender of the OIC argument.

[10] I take these from Peter Vranas ([2007]). See his paper for extensive citations of authors who have given such arguments for OIC principles.

good with respect to (OIC), then they are equally good with respect to (OIVC). Thus, if you accept (OIC) for these traditional reasons, you should accept (OIVC).

Second, much of the plausibility of (OIC) comes from considerations about what is available to the agent under consideration. We noted this above, when clarifying the argument. Consider again the chess example: though I *can*, in one sense, beat Karpov, beating Karpov is not an action that is available to me. It is not an action that I can bring about if I so desire. So, it is not an action that I can do, relevant to (OIC). Put another way, something that I cannot reliably bring about is not something I can *do*, but rather something that *happens to* me. (OIC) is concerned with the actions that are available to the agent and that the agent can do, which is why (OIC) is stated in terms of what the the agent can reliably bring about.

Given that (OIC) rests on this notion of availability, however, it seems that (OIVC) is strongly supported. Something over which an agent has no voluntary control is something that might happen to the agent, but certainly is not something available to the agent in the practical sense that matters for obligation. For instance, I have no voluntary control over digesting my food. My food is nevertheless reliably digested. But it is a mistake to think that digesting my food is an option that is available to me. The lack of voluntary control over it makes it no more an available action to me than beating Karpov at chess. After all, digesting my food is not something that I *do*, but rather something that happens *to me*. Similarly, if an agent has no voluntary control over conforming to a principle then conformance is not something the agent *does*, but rather something that *happens to* the agent. So, a lack of voluntary control implies a lack of practical availability. But this lack of practical availability is exactly what prohibits an

action from being obligatory according to (OIC). Thus, this notion of availability that features prominently in (OIC) also tells in favor of (OIVC).

Finally, one can argue that it is implausible to deny (OIVC) while maintaining (OIC) because (OIVC) is *conceptually prior* to (OIC). What is clear, the thought goes, is the truth of something like (OIVC). After all, it is plausible that we are under no obligation to beat our hearts, not because we *cannot* beat them but rather because beating our hearts is not under our voluntary control. We establish (OIVC) first. Then, since (OIVC) entails (OIC), (OIC) receives support from this. If one affirms (OIC) while denying (OIVC), things look backwards.

The basic moral is that it is unmotivated to uphold (OIC)—in the specific form that it is needed for the OIC argument against BE—while denying (OIVC).[11] If this is right, then one who upholds the OIC argument against BE must reject (V2) of the Anti-Voluntarist Argument. Thus, such a person must embrace doxastic voluntarism. This in itself is surprising. I will next argue that it is also implausible.

1.5.2 Premise (V2)

To deny (V2), one must claim that we sometimes have control over our doxastic attitudes. But, realistically, one must do more than this. (OIVC) together with the OIC defender's claim that rationality gives us obligations, entails that a theory of rationality *only* concerns those doxastic attitudes over which we have voluntary control. So,

[11] Note that there are OIC principles that don't have all these connections to (OIVC). As noted above, it is (in some sense) possible for a human to conform with *PROB* and *COND*, just as it is (in some sense) possible for Adam to predict the outcome of the dice rolls. But there are no procedures that are available or non-lucky that bring this about, so such possibilities do not satisfy (OIC), which is behind (2′). This suggests that a different OIC principle, which counts *PROB*/*COND*-conformance and Adam's prediction as possible, will be easier to accept while rejecting (OIVC). As we saw, however, this is not the kind of principle needed by the defender of the OIC argument.

although strictly speaking the Anti-Voluntarist Argument is avoided by showing *one* doxastic attitude under voluntary control, the defender of the OIC argument really needs to show a large class of doxastic attitudes that are under voluntary control. If we have no control, then the Anti-Voluntarist Argument is not avoided. If we have very little control, then our theories of rationality are in danger of being severely constrained. If we have robust control, then there is no problem. Since the denial of (V2) is an empirical claim with no clear answer, it is difficult to tell exactly how things stand. However, I will argue that we have reason to think that things are not good. Essentially, then, the argument is that by endorsing the OIC argument against BE, one places a heavy restriction on what theories of rationality can be about in a way that is theoretically undesirable.

First we need to ask whether or not we have voluntary control over our doxastic states. Working through this will give us a more realistic picture of what sort of doxastic voluntarism is plausible. Steven Sloman ([2002]) demarcates two reasoning systems in humans: the associative system and the rule-based system. He claims that the latter is automatic, whereas the former is deliberate or controlled. This sounds like good news if one wants to reject (V2). However, we must look at this claim carefully.

Sloman understands the controlled/automatic distinction as given by Shiffrin, Dumais, & Schneider ([1981]). According to Shiffrin, et. al. ([1981]), a process is automatic when two conditions are met:

> 1: Any process that does not use general, nonspecific processing resources and does not decrease the general, nonspecific processing capacity available for other processes…2: Any process that demands resources in response to external stimulus inputs, regardless of participants' attempts to ignore the distraction. (quoted in (Sloman [2002], p. 393)

Thus, a process is automatic if (1) it doesn't take up general processing resources and (2) you cannot deliberately stop it. Sloman doesn't tell us what a controlled process is, but perhaps we are to understand a controlled process as one that is not-automatic. A process would be not-automatic if it *did* take up general processing resources, and *were* deliberately stoppable. Thus, if the rule-based reasoning system is not-automatic, we would have a certain kind of voluntary control: control over the disactivation of this reasoning system.

But this is not all good news for the defender of the OIC argument. Disactivation of a reasoning system could be voluntary without the *way* of reasoning being voluntary. Unless the defender of the OIC argument wants to deny that theories of rationality can be about *ways* of reasoning, he needs it to be the case that our *ways* of reasoning are under voluntary control. Sloman's work does not support this.[12]

If our reasoning system is simply not-automatic, this would limit our theories of rationality too much. But perhaps there is more to being a controlled process than simply being not-automatic. In an earlier paper Schneider & Shiffrin ([1977]) tell us more about controlled processes:

> A *controlled process* is a temporary sequence of nodes activated under control of, and through attention by, the subject…Controlled processes are therefore tightly capacity-limited, but the cost of capacity limitations is balanced by the benefits deriving from the ease with which such processes may be set up, altered, and applied in novel situations for which automatic sequences have never been learned. (Schneider & Shiffrin [1977], p. 2)

[12] In addition, Sloman's distinction is controversial, and the evidence he gives for the claim that the rule-based system is controlled is that conscious attention to the processes performed by the system is required. Obviously, attention to a process does not mean voluntary control over the process. This is not a criticism of Sloman. As far as I can tell he does not claim that the processes are under voluntary control, only that they are *controlled*.

If we assume that there is a reasoning system that is a controlled process in this sense, then things look better for the defender of the OIC argument. Assuming that control over a process means more than just attention to the process, if we had this kind of control over our reasoning system, it looks like we could set up new processes, alter existing ones, and apply these processes in novel situations. This looks more promising.

However, caution is required. Schneider & Shiffrin are concerned with how we learn new arbitrary correlations between objects (in particular, letters of the alphabet) and then store this information in our memory. It is not clear that what they say is relevant to the question of doxastic voluntarism. If we look at the subjects in Schneider & Shiffrin's study, the kind of control that they have over cognitive processes depends on putting themselves through memorization training regimens. Subjects study certain groupings of target letters, and are then able to recognize these target groupings more quickly. Thus, even if there is some reasoning system in us that is a controlled process in this sense, the kind of control would concern the training regimens that the agents are able to put themselves through. This kind of control appears to be *indirect* rather than *direct*: subjects can *indirectly* control their ways of reasoning by *directly* controlling which training regimens to undergo.

Given this, it seems that, at best, we have voluntary doxastic control of a restricted kind. Since we are concerned with accounts of epistemic rationality, the question for us is how this kind of control could bear on the adoption of reasoning strategies. The best case for the doxastic voluntarist is that we are able to voluntarily put ourselves through training regimens that store in memory certain ways of forming beliefs. Perhaps, then, subjects could learn to automatically change and manage their

beliefs in certain ways. Now, it is unlikely that any kind of perfectly *general* rule—such as the rule that one must be coherent—could be voluntarily adopted. But one might be able to learn more specific rules that govern belief-formation in a specific domain. For example, imagine an avid gambler who knows the gambler's fallacy is a fallacy, but can't shake the belief that, after a long run of black, red is much more likely than 50%. With enough practice, perhaps the gambler could train himself *not* to have a high degree of belief in red in response to long runs of black.

Claiming that we have *this* kind of voluntary control over our beliefs is in agreement with Alston ([1989]), who argues that we do not have *direct* voluntary control over our belief states, in the sense that one has direct voluntary control over moving one's arm. If we have voluntary control over a belief state, according to Alston, it is *indirect* control. Alston actually distinguishes two types of control that I have called 'indirect control.' The first is *long-range control*: "…the capacity to bring about a state of affairs, C, by doing something (usually a number of different things) repeatedly over a considerable period of time…" ([1989], p. 134). The second is *indirect voluntary influence*. Essentially, this is voluntary control over things that influence, though don't bring about, a certain doxastic state. Alston gives the following examples: "…I have voluntary control over whether, and how long I consider the matter, look for relevant evidence or reasons, reflect on a particular argument, seek input from other people, search my memory for analogous cases, and so on." ([1989], p. 138) These things are not *guaranteed* to change my doxastic state, but they certainly influence that state.

Once we think about indirect control (whether long-range or indirect influence), we notice that there are many ways of voluntarily affecting our belief states and our

ways of forming beliefs. For instance, when faced with a problem about percentages I can voluntarily think about it in terms of set-membership. This may reliably result in me holding different beliefs about the issue than I would have had I continued to think in terms of percentages.[13] Or, when I am about to perform a difficult calculation, I might decide to go fetch a calculator. Again, this will reliably result in me holding different beliefs than I would have held. These are both things over which I have voluntary control. Thus, we might say that humans *do* have voluntary control over their doxastic states and their ways of reasoning in virtue of these kinds of indirect control. Denying (V2), perhaps, doesn't look so bad.

However, things are worse than they seem. First, distinguish between strong and weak (indirect) control. Say that strong (indirect) control over P is the ability to do something (that isn't P itself) that reliably brings about P. This corresponds, roughly, to Alston's *long-range control*. Weak (indirect) control over P is the ability to do something (that isn't P itself) that raises the chance that P comes about and isn't strong control. This corresponds, roughly, to Alston's *indirect voluntary influence*. The kind of control that it looks like we have over our beliefs is weak control. We can influence how we reason and how we form our beliefs, but we cannot reliably bring about changes of reasoning in a strong sense. Perhaps, for instance, by taking more courses in statistics, one weakly influences one's beliefs so that they better conform with *PROB*. This is an example of weak control.

This allows one to deny (V2) if one holds that weak control is all that is necessary for the truth of doxastic voluntarism. Thus, perhaps the defender of the OIC

[13] See, for instance, Gigerenzer, ([2000]) and Lagnado & Sloman ([2004], p. 171).

argument can deny (V2) and escape the anti-voluntarist argument. This, however, is mistaken. For once we distinguish between strong and weak control, we must distinguish between two versions of the anti-voluntarist argument: one that appeals to strong control and one that appeals to weak control. Accepting doxastic voluntarism in the weak sense allows one to deny (V2) of the weak control version of the argument. But commitment to (OIC), the principle that features in the OIC argument, makes plausible an OIVC principle formulated in terms of strong voluntary control:

(OIVC-strong): If humans have an epistemic obligation to believe in some way, then humans have strong indirect control over believing that way.

This is because there *are* things that we can do which will indirectly make us better at conforming to *PROB* and *COND*. We can, for instance, take more courses in statistics. But the defender of the OIC argument maintains that this is not enough. (OIC) is violated even if we can indirectly influence our *PROB* and *COND* behavior. But then commitment to (OIC) carries with it commitment to (OIVC-strong). Thus, it is the Anti-Voluntarist Argument that we get using (OIVC-strong) to support (V1) to which the defender of the OIC argument must respond. But this argument does not permit one to reject (V2) by appeal to weak control. To respond to this argument, the defender of the OIC argument must claim that we have strong indirect control over our beliefs.

This position is problematic. First, it is not clear how much strong indirect control we have over our belief states or our ways of forming beliefs. But, more importantly, (OIVC-strong) significantly reduces the range of theories of epistemic rationality that are open to consideration. According to (OIVC-strong), we only have epistemic obligations to believe in certain ways if we have strong indirect influence over believing in those ways. Since the defender of the OIC argument thinks that all

claims about epistemic rationality entail corresponding epistemic obligations, we would be entitled to claims about epistemic rationality only for those aspects of our doxastic lives over which we have strong indirect control. Just how much of our doxastic lives is subject to such control is of course an empirical question. But I think it is an undesirable consequence of this view that theories of rationality are restricted in such a way. I do not mean to claim that a theory that restricted itself in such a way is undesirable. Indeed, an account of what parts of our epistemic lives we have voluntary control over seems to be of general and philosophical interest. What I think is undesirable is the claim that theories of rationality are *only* about that. By claiming that rationality gives epistemic obligations, and that there is a conceptual entailment between epistemic obligation and possibility, however, the defender of the OIC argument is committed to just such a claim.

Consider an analogy: A theory that concerned itself with what training regimens one can undergo to improve one's physical body is interesting. Nevertheless, a theory of physical excellence need not *only* concern itself with those things one can reliably bring about via training regimens. It might also apply to things about the physical body itself, things over which the agent has no strong indirect control.

The defender of the OIC argument faces a dilemma. He must reject either (V1) or (V2). Denying (V2) is implausible and forces a restriction on theorizing about rationality. However, (OIC), which is needed for the OIC argument, prohibits one from denying (V1). Because of this, we can reject premise (2′) of the OIC argument against BE. It should be rejected, because it leads us to untenable positions when we consider the Anti-Voluntarist Argument. This is an attractive response to the OIC argument

because the status of (2′) does not depend on a stalemate of intuitions about correct OIC principles.

Although this response to the OIC argument has that good feature, it is also lacking in at least two ways. First, though it does provide a response to the OIC argument, it doesn't at all address the general issue of how an agent's capacities should bear on epistemological theorizing. That doesn't invalidate the response, but it does leave us with more to do. Second, this response depends critically on a certain sort of theoretical desiderata: that our theories of rationality are not too much limited by considerations of voluntary control. But, of course, it is open to the defender of the OIC argument to claim that this kind of limitation is just what is called for. For these two reasons, I next consider a different sort of response to the OIC argument. Unlike the response considered above, this second response is more prominent in the literature.

1.6 Misunderstanding Bayesian Epistemology

I turn now to the second response to the OIC argument. The main claim will be that the OIC argument's success rests on misunderstanding BE in an important way.

There are two well-known ways to understand a non-descriptive theory in a certain domain: as either an evaluative theory or as a normative theory. If a non-descriptive theory is evaluative, then though it sets certain standards, there are no obligations issued by that theory. Consider: we can have an evaluative theory of the weather, which tells us what sort of weather is good and bad, without claiming that the weather is under any obligation to be a certain way. On the other hand, if a non-descriptive theory is normative, then it is thought to place us under certain obligations. Consider: the traffic laws appear to be a normative rules that place obligations on us.

Although the traffic laws do provide a way to evaluate action, they do not *only* evaluate action. They also place obligations on me: I ought not break the traffic laws.

It is plausible that non-descriptive epistemological theories permit of this sort of division. On the one hand there are those epistemological theories concerned solely with evaluation, and on the other there are those that evaluate *and* place normative constraints on agents. It is only the latter, however, that give epistemic obligations. Now, if there is this sort of division of types of non-descriptive epistemological theories, then there is a very clean response to the OIC argument against BE. Early in this paper I characterized the Bayesian account of rationality as a normative theory, in contrast to a descriptive theory. But we can now see that such a characterization may have been imprecise. I should have said that the Bayesian account of rationality is a *non-descriptive* theory, which is consistent with it being either evaluative or normative. The response to the OIC argument, then, is to claim that the accounts of rationality we get from BE are *evaluative*, rather than *normative*. Since evaluative theories give no obligations, premise (1) of the OIC argument is false.

Some, however, may feel that it isn't quite right to say that an epistemological theory could be *wholly* evaluative. Thus, seeing the Bayesian account of rationality as an evaluative theory that gives no obligations would be unacceptable. However, Matthew Chrisman ([2008]) has recently proposed a view of epistemic obligation that allows us to draw a distinction something like the evaluative/normative distinction and yet maintain that epistemological theories always give obligations. Adopting his account, we can formulate the same kind of response to the OIC argument, even if one doesn't want to claim that BE gives evaluative theories.

Chrisman draws a distinction between *ought-to-be* obligations and *ought-to-do* obligations. The former attach to states of affairs, the latter to actions. For example, we might say that it ought to be that children can tie their shoes (ought-to-be obligation) and that you therefore ought to teach them (ought-to-do obligation). Chrisman then argues that ought-to-be obligations do not imply voluntary control. For example, it may be true that it ought *to be* that Big Ben keeps good time, without supposing that Big Ben has voluntary control over whether or not it keeps good time. Chrisman further urges that our epistemic obligations are ought-to-be obligations. Since these sorts of obligations do not imply voluntary control, he uses this to escape the Anti-Voluntarist Argument.

This distinction between ought-to-be obligation and ought-to-do obligation roughly approximates the distinction between a normative and evaluative theories. Once we draw the ought-to-be/ought-to-do distinction, however, we need to specify which kind of obligation features in (OIC). Recall, (OIC) says that if you ought to conform to a principle, then you have the capacity to reliably bring about such conformance. It seems clear that *if* there is a relation between obligation and this sort of capacity to reliably bring something about, it applies primarily to ought-to-do obligations.[14] This is in agreement with the spirit of Chrisman's proposal where he argues that ought-to-be obligations do not imply voluntary control. It is also in agreement with Vranas ([2007]) who defends a version of an OIC principle only with respect to obligations to *do* something.

[14] One may think that ought-to-be obligations *do* have implications about what is possible. My claim is that if they do, it will be via some connection to ought-to-do obligation. This is also in agreement with Chrisman's general thesis.

In summary, we can understand an epistemic theory as either evaluative or normative. If the former, then that theory places no obligations on us. Thus, premise (1) of the OIC argument is false. On the other hand, we might understand an epistemic theory as normative. If so, then it looks as though that theory will place obligations on us. However, we must determine whether or not these obligations are of the ought-to-be sort or of the ought-to-do sort. If the former, then premise (2) of the OIC argument is false: ought-to-be obligations are not subject to (OIC).[15] Thus, the advocate of BE has a second response to the OIC argument: it misconstrues the BE as being *directly* about what agents ought to do. BE, however, is properly understood as giving evaluative theories or theories about what ought to be.[16] Thus, premise (1) of the argument is false.

It is worth noting that this response is related to the Anti-Voluntarist response. In both cases, the problem for the OIC defender is a restricted view of theories of rationality. The first response showed that the OIC defenders are committed to theories of rationality that only speak to things over which agents have some measure of voluntary control. This second response shows that the OIC defenders are committed to understanding BE as only giving us normative theories issuing in ought-to-do obligations.

[15] It is important to be clear about this. The claim is *not* that there are no relations between obligations about what ought to be and obligations about what to do or between evaluative claims about what is good and what we thus should do. The problem is rather that there are very many relations that might hold between these things, and that a very particular kind of relation is needed for the OIC argument to go through.

[16] For instance, Christensen ([2004], ch. 5.

1.7 Understanding Bayesian Epistemology

Although claiming that BE is evaluative[17] *does* get one out of the OIC argument, this response suffers from one of the worries with the Anti-Voluntarist response: there is still the sense that we haven't fully addressed the issue of how an agent's capacities bear on epistemological theory. For note that one can easily have an evaluative epistemic theory that *does* take account of an agent's abilities. We can, for instance, evaluate how well some agent solves a problem, given the finite computational abilities that the agent possesses. From the fact that an epistemic theory is evaluative, nothing follows about whether consideration of the capacities of the agents is relevant.

This is not a criticism of the response to the OIC argument. That response is a good one. But if one thinks that evaluative epistemic theories should be sensitive to the limitations of agents, then the victory over the OIC argument is a victory of letter, not of spirit. For the objector to BE could come back and say that although the OIC argument has been disarmed by claiming that BE gives evaluative theories, even evaluative theories can and should take account of the capacities of agents.

For this reason, the claim that BE gives us evaluative epistemic theories seems best understood not just as a particular way of escaping the OIC argument, but rather as part of a broader claim about how to understand BE. In particular, the suggestion is that not only is BE evaluative, but it is evaluative in a special sort of way that renders its inability to be followed by actual agents irrelevant. This sentiment seems to be expressed by Christensen ([2004]):

[17] In the previous section I said that Bayesian Epistemology should be seen as evaluative, or as giving ought-to-be obligations. In this section, I'll shorten this for brevity and just say that BE should be seen as evaluative.

> There is nothing mysterious about evaluative concepts whose application is not directly constrained by human limitations, even if the evaluations apply to distinctively human activities. To look at just one example, consider goodness of chess play. We can see right away that chess judgments typically abstract from the psychological limitations of individual players… Moreover, our fundamental metric of goodness for chess play flows from an ideal that is not even limited by *general* human psychological constraints. True, our ordinary quality judgments about chess players are expressed in terms that are relativized to general human capacities… But underlying all of these relativized judgments is an absolute scale that makes no reference at all to human cognitive limitations…There are some arenas in which perfection is humanly attainable, and some in which it is not. The considerations above merely suggest that, in this respect, rationality is more like chess than it is like tic-tac-toe. (pp. 163-4)

Kaplan ([1996]) can also be seen as giving voice to this basic idea. Kaplan defends a view he calls 'Modest Probabilism'. In developing this view, he draws a distinction between a rational person and a rational state of opinion. He writes:

> Being a regulative ideal, Modest Probabilism is not to be understood as giving you a set of marching orders which you can violate only at the cost of being classified as irrational. Modest Probabilism is, rather, to be understood as putting forth a standard by which to judge the cogency of a state of opinion – a standard according to which any violation of Modest Probabilism opens your state of opinion to legitimate criticism. ([1996], p. 37)

Kaplan's Modest Probabilism, which is a theory in BE, is meant to apply to states of opinion, not persons. This is a way of evading certain worries that arise concerning computational ability.[18] I suggest, then, that understanding theories in BE in accordance with Kaplan and Christensen does not simply provide a response to the OIC argument, but also offers a unified way of dealing with a wide range of general complaints about the computational limitations of epistemic agents.

---

[18] In his ([1996]), Kaplan writes about why he rejects a certain principle he calls 'Immodest Connectedness':

> It is important to notice that my reason for rejecting as falsely precise Immodest Connectedness's demand that you place a monetary value on each well-mannered state of affairs is not what one might have expected. **It is not that this demand is not humanly satisfiable**. For if *that* were all that was wrong, the demand might still play a useful role as a regulative ideal… (p. 29, my boldface)

According to this view, BE is not just to be understood as giving *evaluative* theories, but as giving a particular kinds of evaluative theories. It gives us theories whose aim is *not* to provide an account of *how*, given the resources that we actually have, we should go about getting certain beliefs, or what procedures we should follow. We could call such theories 'non-procedural theories'. Thus, BE gives us evaluative, non-procedural theories. There is precedent for understanding evaluative theories in this way.

For instance, understood in this way, BE is similar to what some think the theory of Utilitarianism provides us with in ethics.[19] Utilitarianism, according to this view, gives us a standard of rightness. Thus conceived it doesn't tell one how to do the right thing, it merely tells one what the right thing is. Similarly, BE doesn't tell one how to do the epistemically correct thing, it merely tells one what the correct thing is.

This sort of understanding of BE also fits well with a distinction that Herbert Simon ([1986]) draws with respect to theories of rational economic behavior. Simon draws a distinction between what he calls *substantive* and *procedural* theories of rationality:

> Behavior is substantively rational when it is appropriate to the achievement of given goals within the limits imposed by given conditions and constraints… Behavior is procedurally rational when it is the outcome of appropriate deliberation. Its procedural rationality depends on the process that generated it. ([1986], pp. 130-1)

He later writes that procedural rationality is concerned "with the actual processes of cognition, and with the limits on the human organism that give those processes their peculiar character." ([1986], p. 147)

The implication of this claim is that BE is a regulative ideal and so human limitations are not objections to it.

Simon draws this disinction while discussing theories of rationality in decision situations, and thus is considering practical rationality. In standard decision theory, the rational decision is the one that is optimal, where optimization is understood in terms of EUMAX. If this optimal option is selected, we say that the agent made the rational decision. The theory is silent, however, about *how* an agent manages to reach this decision, or even if it is possible for the agent to reach such a decision. This, according to Simon, is a substantive theory of practical rationality. Simon argues, however, that it is often useful and practical to characterize a rational decision as the one that resulted from the employment of a rational decision procedure. A theory that attempts to capture this aspect of rational decision-making is a procedural theory of practical rationality.

To say that BE gives us anti-procedural theories, then, is to say that it is *not* interested in the sorts of questions pursued by procedural theories as understood by Simon. A procedural theory would attempt to provide an account of *how*, given the resources that we actually have, we should go about getting certain belief states. But this is not the concern of BE.[20]

It is important to note that Simon's procedural/substantive distinction is not perfectly analogous to the distinction I wish to draw. Simon's distinction seems to imply that any theory that references processes is going to be a procedural theory. But

---

[19] For instance, see Bales ([1971]).

[20] Frederic Laville, in his ([2000]) builds on Simon's ideas about substantive and procedural rationality. He notes that appealing to this distinction can allow us to understand a dispute that arises surrounding how best to understand optimization theory (that is, a theory of choice structured around optimizing). First, we can understand optimization theory in what he calls an instrumentalist way. According to this view, optimization theory is a substantive theory: the point of the theory is *not* that people actually calculate things according to optimizing formulas. Second, we can understand optimization theory in a realist way. According to this view, optimization theory is a procedural theory: if you don't calculate according to the optimizing formulas, then your behavior doesn't count as rational. I am maintaining that there are two corresponding ways of seeing theories of *epistemic* rationality, and that BE is best seen as giving theories of the first sort.

this is not what is intended here. An anti-procedural theory could still reference processes of belief formation in various ways. For instance, consider a process-reliabilist account of justification according to which a belief is justified just in case it is produced by a reliable process. Is this a procedural or an anti-procedural theory? It might seem to be a procedural one, for process-reliabilism references processes of belief formation in its statement of what makes for a justified belief. Further, one could say, it does give a procedure that should be followed: only hold beliefs that result from reliable processes. Nevertheless, I think it is best thought of as an anti-procedural theory, for it doesn't tell one how to determine which beliefs are the result of reliable processes. So it doesn't give one a procedure for getting the "good" beliefs. That is, it tells one what the good states of belief are, but not how to go about getting into those states of belief. In a similar way, BE *could* be seen as giving us a procedures to follow: only hold beliefs that are sanctioned by *COND* and *PROB*. This *is* a procedure, but it is one that we can't hope to follow, nor do those who propose theories in BE intend people to follow them. Accordingly, BE gives us anti-procedural theories.

Consider an analogy. An anti-procedural theory does something like specify a function, whereas a procedural theory does something like specify an algorithm. Strictly speaking, a function isn't the sort of thing that you assess for computability. Algorithms are the sorts of things that are assessed in this way, not functions. Now, sometimes we understand the question: 'how much computing power does that function require?' to be asking for the computing power for the least demanding algorithm that will compute the function. But strictly speaking it is algorithms that have computational properties, not functions.

To illustrate, let's assume that the results of a certain function have some property that we find desirable or good. From the fact that the results of this function have this good property, nothing follows about what should be computed. Before we can talk about computation, we must specify an algorithm. Now, perhaps whenever there is a function with this good property, we should implement an algorithm that can, with certainty, compute this function. But this is not the only possible view. Perhaps we should implement an algorithm that can, with high probability, compute the function. Or, perhaps we should implement an algorithm that can approximate the function. Issues about computability come in at the level of algorithms, not functions.

There is room in epistemology for anti-procedural theories, which, on this analogy, set themselves the task of specifying functions, not algorithms. Considerations of abilities need not be relevant to such theories. In this way, BE gives us anti-procedural theories. As Christensen implies, BE is meant to provide an absolute standard for goodness of reasoning that is not limited by cognitive ability. This not only gives BE a response to OIC arguments (by implying that it is evaluative). It also provides a unified way to answer questions about how the abilities of epistemic agents bear on epistemic theorizing for BE.

1.8 Conclusion

I began by considering and clarifying the OIC argument against Bayesian Epistemology. I then offered two responses. The first showed that the success of the argument depends on a certain conception of theories of rationality that is theoretically limiting. In particular, defending the OIC argument runs one into trouble with doxastic volunatarism. The second response showed that the success of the OIC argument

depends on a certain way of understanding what it is that Bayesian theories are attempting to do. By clarifying the role of such theories, the OIC argument can be avoided. The main lesson, then, is that the OIC argument against Bayesian Epistemology can be avoided. But that doing this depends on understanding Bayesian Epistemology in a certain way. We should understand it as giving us evaluative anti-procedural theories. This allows the Bayesian theorist to avoid OIC considerations in particular, and general considerations concerning the limited abilities of epistemic agents.

CHAPTER 2

EVIDENCE: INTERNAL AND EXTERNAL

2.1 Introduction

Bayesian Epistemology (BE) is a framework for constructing epistemological theories. There are several themes that unify these theories. First, agents are represented as having partial beliefs, which are to meet certain synchronic constraints. Second, these beliefs are to change in a certain way, and so meet certain diachronic constraints. Third, this kind of belief change is to be prompted by the receipt of evidence. There are further considerations relevant to practical decision-making, but I focus here on the epistemic aspects of BE. Thus, we have three important aspects to Bayesian Epistemology:

i. Synchronic constraint on partial beliefs (*PROB*).[1]

ii. Diachronic constraint on partial beliefs (*COND*).

iii. Belief change initiated by receipt of evidence.

In the last chapter, I focused on i and ii, and noted that by understanding Bayesian Epistemology in a certain way, we are able to dismiss objections to *PROB* and *COND* based on computational considerations. In this chapter, I will focus on iii, and what should be said about evidence.

There are two main goals for this chapter. First, I want to explain why it is important to the project of Bayesian Epistemology to have an account of what it is to have evidence. Second, I want to discuss what kind of account of evidence we should

[1] Sometimes *PROB* is taken not to be an epistemological *constraint*, but rather taken as a restriction on the kinds of agents Bayesian Epistemology applies to: agents whose beliefs are probabilistically coherent. See, for instance, Meacham ([*forthcoming*]). However, as the proponderance of arguments in favor of probabilistically coherent beliefs attest, it is common to understand *PROB* as a substantive constraint.

be looking for. I'll particularly focus on whether or not the account of evidence for Bayesian Epistemology should be an internal one or an external one. As a rough characterization, an externalist account of evidence allows that internally identical agents can have distinct evidence. By specifying 'internally identical' in various ways, one gets different forms of externalism about evidence.

One might think that Bayesian Epistemology is firmly wed to internalist ideas, and so that an externalist account of evidence is incongruous with the general thrust of BE. Since in subsequent chapters I wil propose and defend an account where features external to the agent determine what evidence that agent has, it is important for me to disarm this thought. The bulk of this chapter, then, is defensive, attempting to dissuade one from the thought that Bayesian Epistemology is necessarily tied to internalist epistemology. I will also suggest that BE, understood from Chapter 1 as evaluative and anti-procedural, fits nicely with an externalist account of evidence and so gives some reason to pursue such an account.

Recall what it is to understand Bayesian Epistemology in an evaluative, anti-procedural way. An evaluative theory is a theory that does not issue obligations. It merely sets a standard against which performance can be evaluated. An anti-procedural epistemological theory is a theory whose aim is not to provide an account of *how*, given the resources they actually have, agents should go about getting certain beliefs. So, an evaluative, anti-procedural Bayesian Epistemology gives theories that set a standard against which epistemic performance can be evaluated that is not concerned with specifying a procedure for coming to have certain beliefs. In this chapter I will argue

that if we understand Bayesian Epistemology in this way, then many of the motivations for internalism about evidence disappear.

Here is a roadmap for the chapter. I'll first point out why we must say something substantive about evidence. I'll make this point by considering a very deflationary notion of evidence, and seeing how it goes wrong. Then, I'll argue that external and internal accounts of evidence both have *prima facie* considerations in their favor. After this I'll consider two main ways to understand internalism—*first* in terms of access, and *second* in terms of guidance—and argue that with internalism so-understood principles of BE actually tells against internalism. Finally, I'll consider a different way of understanding internalism, according to which Bayesian Epistemology appears to be more internalist. However, I will argue that this is still no barrier to pursuing an externalist extension of the theory.

Note, however, what I won't be doing: I won't be saying much about the details of an externalist account of evidence, or how such an account would fit in with Bayesian Epistemology. This is a pressing and interesting question, but I will not address that question yet.[2] Here I set myself the more modest goal of arguing that *some* such externalist account of evidence is worthy of pursuit.

2.2 Why We Need an Account of Having Evidence

There are many things that one might be interested in when one is interested in evidence. To clarify my project, consider two specific things that one might be interested in. First, one might be interested in the evidence-for relation. If one is interested in this, then one is interested in saying when a certain piece of information is

evidence *for* something else. Consider an example (due to Sherrilyn Roush [2006]) to make this clearer. Let's say that in 2002, a shipment of metal tubes is intercepted on its way to Iraq. An important question is whether or not this shipment provides evidence for the proposition that Iraq has a WMD program. One might think that to decide whether or not this is evidence for that proposition, we should look at the probability of the hypothesis on its own in comparison to the probability of the hypothesis conditional on the evidence. That is, we need to look at the ratio: P(H|E)/P(H). If the ratio is 1, then the information is evidentially irrelevant, if the ratio is less than 1, then the information is evidence *against* the hypothesis, and if the ratio is positive, then the information is evidence *for* the hypothesis. Of course, there are many other proposals that can and have been made about this question.[3] Indeed, Bayesian epistemologists and Bayesian philosophers of science have a lot to say about the evidence-for relation. The debate concerning different measures of confirmation—whether the difference ratio, or the likelihood ratio, or the odds ratio, etc.—is a debate about the evidence-for relation. These are indeed important and interesting debates. Much less attended to, however, is the question of when something is to count as *evidence* for an agent.[4] In what follows I will focus on *The Evidence Question*: When and under what conditions does an agent have some evidence?

I claim that an adequate Bayesian Epistemology needs to say something about what kinds of things are to count as evidence for an agent, and thus must answer The Evidence Question. It is all very well to know when some proposition is evidence *for*

---

[2] I do address this question in Chapter 3.
[3] For more details about such proposals, see Fitelson ([2001]).
[4] Williamson ([2000], ch. 8) is a notable exception. Silins' ([2005]) and Neta ([2008]) have recently added to this discussion. See also, Feldman ([1988]) and Maher ([1996]).

some other proposition. But if the first proposition is not part of your evidence, this gives you no reason to believe the second.

Now, one might grant that *something* must be said to answer The Evidence Question, but insist that what must be said is very minimal. In Howson & Urbach's *Scientific Reasoning* ([1993]) a deflationary account of what it is to have evidence is suggested:

> When your degree of belief in *e* goes to 1, but no stronger proposition also acquires probability 1, set $P'(a) = P(a|e)$ for all *a* in the domain of *P*, where *P* is your probability function immediately prior to the change. ([1993], p. 99)

The suggested account is that one's new evidence at some time is whatever one has become certain of at that time. Thus, one's total evidence at a time consists of all those propositions one is certain of at that time. Though this does give an account that specifies what it is for an agent to have evidence, it does so in a way that makes the question seem epistemologically unimportant. Peter Milne, for instance, approvingly cites Howson & Urbach as follows:

> We do not enquire as to why the agent comes to assign full belief to *e* in the interval from *t0* to *t1*. This change is, in Howson and Urbach's felicitous phrase, *exogenous* (1993: 106). Experience may prompt the change, even bring it about causally, perhaps, but the details of the story of how that comes about are of no moment. ([2003], p. 283)[5]

[5] In Howson & Urbach's first addition of *Scientific Reasoning* ([1989]) this idea is present, but somewhat muted. For instance, they write: "…the data *e* appearing in the conditional probabilities are given exogenously: *e* is, for want of a better word, simply 'known'." ([1989], p. 285) The occurrence of the word 'known' suggests something much less deflationary than indicated above in the text, but they do not seem to mean anything substantive by the term 'known'. Rather, it seems to refer to simply full confidence. Later they write: "All that the ascription of probability one to *e* entails, in our and Levi's view, is that the agent takes *e* to be true in the light of his current experience." ([1989], p. 288) Again, this seems somewhat less deflationary than I have indicated above, since they add 'in the light of his current experience.' But this reference to experience does no work for Howson & Urbach's claims about evidence. Rather, it is the fact that the agent takes *e* to be true, that makes it evidence for that agent. This is even clearer in the second edition of *Scientific Reasoning* ([1993]): "Let us call changes of belief which are not conditionalization-changes *exogenous*. So the exogenous change from $P(e) = p$ to $P'(e) = 1$ is required for Bayesian conditionalization to determine the distribution $P'(c) = P(c|e)$ over all those other propositions *c* which the agent contemplates." ([1993], p. 106).

First, note the truth in this claim. It surely isn't the case that *every* Bayesian Epistemologist must say something about what evidence is. It is surely acceptable to focus on one aspect of an epistemological theory, rather than another. But the idea that what it is to have evidence is not epistemologically important, is an idea that should be resisted. I will attempt to show this with examples.

2.2.1 Case 1

According to the standard Bayesian picture, rational belief change occurs via conditionalization. In its most basic form, conditionalization states:

COND: $cr_{new}(\bullet) = cr_{old}(\bullet|E)$

In words this says that your credence function once you've learned E (and nothing else) should be equal to your credence function conditional on E before you learned E. A different, but sometimes more useful way of characterizing conditionalization makes reference to epistemic moments.

COND: $cr_{t1}(\bullet) = cr_{t0}(\bullet|E_{t1})$

In words this says that your credence function at epistemic moment t1 should be equal to your credence function at t0 conditional on the evidence that you received at t1.

Assume, then, that at t0, a rational detective is undecided about the truth of both G and M, where G = "the glove is OJ's" and M = "OJ is the murderer." Further, the detective's conditional credence in M given G is high. Then, at t1, the detective becomes certain of evidence G and adjusts M accordingly. Here is what the whole process looks like, according to the standard picture:

t0: $cr(G) = ½$ $cr(M \wedge G) \approx cr(G)$ $cr(M) = ½$

t1: $cr(G) = 1$ $cr(M \wedge G) \approx cr(G)$ $cr(M) \approx 1$

When this process is described as I just did, it sounds like a paradigm case of updating beliefs by conditionalization. The agent receives some evidence, *G*, and this evidence effects the rest of the agent's beliefs accordingly. But this canonical way of describing the formalism is not the only thing that one can say consistent with that formalism. For all the formal picture tells us, the agent could have come to have a high credence in *M*, and on account of *this*, come to be fully confident in *G*.

We have, then, two very different scenarios, both consistent with the formal account provided by *COND*. The first scenario is one where the agent comes to believe that *G*, and *because of this*, comes to believe that *M*. The second scenario is one where the agent comes to believe that *M*, and *because of this*, come to be fully confident that *G*. These are epistemically different scenarios. In the first, for example, the agent comes to believe some fact about a suspect's guilt in light of some evidence. In the second, the agent come to believe that a suspect is guilty , and then on *that* basis comes to believe there is evidence supporting that conclusion. The latter sequence of belief states is clearly epistemically inappropriate.

To say something about *why* this sequence is inappropriate, we must say something about what evidence is that goes beyond the deflationary account. For in both cases *G* is given full credence, and so in both cases it is counted as evidence. Presumably the thing to say about this is that it is implicit in *COND*, that belief change is spurred on by the receipt of evidence. In the case where the detective believes *M*, and then because of this, believes *G*, belief change is not spurred on by the receipt of evidence. After all, what is wrong with the deviant case is that *G* is being treated as evidence, when it really isn't. That is, the detective is treating *G* as evidence by altering

her beliefs so that it appears as though *G* has been conditionalized on, when in fact *G* is not evidence at all. But we cannot say this if *all* we say is that evidence consists of those propositions with credence 1.

2.2.2 Case 2

Imagine that we have two ordinary humans, Tom and Pete, who are just about to walk out of their air-conditioned building in New York City. At t0, just before going through the door, Tom's credence function is such that cr_{t0}(NYC is hot) = ½. Then, at t1, after walking outside Tom's credence in this proposition goes to 1, with the rest of his beliefs being properly conditionalized on this information. On the other hand, at t0, Pete's credence function is such that cr_{t0}(Sydney is hot) = ½. Then, at t1, Pete's credence in this proposition goes to 1, with the rest of his beliefs being properly conditionalized on this information. All else being equal, Tom's change of belief is more appropriate than Pete's. Both agents meet the Bayesian requirements of *PROB* and *COND*. However, intuitively, the proposition that Pete conditionalized on is not evidence for Pete, given that he is an ordinary human in New York, while the proposition that Tom conditionalized on is evidence for Tom, given the description of the situation.

Now, it is possible that this verdict is incorrect. If, for instance, we learn that for Pete, cr(I'm in Syndey) is high, then perhaps it is correct to say that the changes in belief were equally rational. But put this possibility aside. In the case where Pete doesn't have a high credence that he is in Sydney, the initial verdict stands. It seems that what differentiates Pete from Tom has to do with a difference in the evidence that they conditionalized on, since by stipulation, they both met the other constraints. But then if

we want to mark a difference between Pete and Tom, we need to say something about what evidence is that goes beyond the deflationary account.

These two cases demonstrate that it is not enough to simply say that one's evidence is whatever receives credence 1. To do so is to run roughshod over epistemic distinctions that appear important. Further, these examples demonstrate the more general point that something needs to be said about what evidence is. It is surely epistemically important to distinguish between propositions that should serve as evidence and those that should not. Since a central feature of Bayesian Epistemology concerns how we update our beliefs due to the evidence that we receive, an account of evidence is relevant to Bayesian Epistemology. If it is true that rationality requires one to respond appropriately to the evidence, then we must have something to say about what evidence is. Otherwise, what prevents us from saying that rationality requires one to respond appropriately to any old thing we like (e.g., the first sentence of every chapter of *Huck Finn*)?

2.3 Internalism and Externalism: Preliminary Ideas

I have completed one of my objectives, which was to argue that Bayesian Epistemology needs to say something about what evidence is that goes beyond the deflationary account. In this next section, I will explore some suggestions that attempt to say something about what evidence is. Recall that my primary motive is to argue in favor of pursuit of an externalist account of evidence, so I am most interested in the difference between internal and external accounts.

A very natural idea goes against externalism about evidence, and holds that we can understand what evidence an agent has purely in terms of internal features of an agent. A simplistic kind of internalism about evidence holds that we can get a good account of what agent S's evidence is, purely by looking at S's doxastic state. Implied by this sort of view is that what is missing in the "bad" cases above are the right sorts of beliefs. For instance, one might propose that if Pete is to conditionalize on some evidence proposition, it must not only be the case that his credence in that proposition changes to 1, but he also must have some sort of belief *about* the evidential status of that proposition. So, let E = "Sydney is hot," and let M = "At t1, E is evidence for me." The suggestion is that E is evidence for Pete, and so Pete can rationally conditionalize on E, only if $cr_{t1}(M) = 1$. This will not work, however, for Pete is launched down a regress. Surely the same question that arose for E also arises for M, and thus M would require some sort of meta-meta-evidential belief.

Alternatively, one might claim that Pete's evidence is M and *E together*. Thus, E isn't his evidence, but rather both M and E. On this version of the proposal Pete faces no regress, but it looks as though Pete can simply come to believe M and E in as arbitrary a way as Pete came to believe E. Thus, we face a sort of explanatory regress: wherever we stop, we are left with no assurance that the beliefs were not arbitrarily adopted. Appealing to the agent's other beliefs in this way does not appear to be the right approach to saying what evidence *is*.[6]

[6] This, of course, isn't to say that having beliefs about one's evidence cannot sometimes play a positive epistemic role, nor is it to say that having beliefs about one's evidence is not sometimes epistemically required. It is simply to say that having beliefs about one's evidence cannot be what *makes* the evidence evidence.

Let us generalize the discussion. The simple internal requirement just discussed is a specific kind of *doxastic account of evidence*, which holds that what S's evidence *is* is purely determined by S's doxastic state. Howson & Urbach's deflationary account of evidence is one kind of doxastic account, since it is purely a feature of an agent's doxastic state which propositions receive credence 1. Cases 1 and 2 showed this to be incorrect. The simple internal requirement we've just looked at is a different version of the doxastic requirement on evidence. It too fails. I think that these two examples are a good indication that any doxastic account of evidence will fail. Indeed, we can give an argument for this: Consider two agents with identical doxastic states. Suppose that both fully believe that they are seeing the sunset over the Grand Tetons. One of the twins really is having an experience as of a sunset over the Grant Tetons, while the other is not. Since they have identical doxastic states, there can be no epistemic differences between them in virtue of their distribution of beliefs. According to doxastic accounts of evidence, these doxastic twins must have the same evidence. But this does not allow us to distinguish the two in the way that we would like: either they both have as evidence that they are seeing the sunset over the Grand Tetons, or neither of them do. But this is implausible.[7]

In response to the failure of the doxastic account, we are pushed to a non-doxastic accounts of evidence. A non-doxastic account says that what an agent's evidence *is* is determined by at least some things that are not features of the agent's

[7] James Joyce ([2004]) suggests (though doesn't argue for in detail) that we might get a good account of what an agent's evidence is by appealing to the notion of resiliency. The idea is that P is evidence to degree n just in case $\mathrm{cr}(P) = n$ and this value is resilient in the sense that for all (or most) other propositions, X, $\mathrm{cr}(P|X) \approx n$. This is a doxastic account of having evidence, in that it appeals to nothing other than the agent's belief state. Though interesting, I think it fails for the same reason that the more simple doxastic accounts fail: nothing guards against an agent who simply adopts, for no reason, a

doxastic state. Since there are non-doxastic but still internal features of epistemic agents, there can be non-doxastic accounts of evidence that are not externalist accounts of evidence. Nevertheless, many non-doxastic accounts of evidence are externalist. For example, one such account might appeal to how an agent's belief was formed to say which beliefs are to be evidence. A different account might appeal to features of the agent's environment to say which facts about the agent's environment constitute the agent's evidence, whether or not they are believed.

Some sort of external account of evidence would seem able to quickly handle Cases 1 and 2. To fix ideas, assume that we adopt an account that says that an agent's reliably formed beliefs are his evidence. Now, in Case 1 there are two different ways in which the agent meets the requirements of *COND*. But the *way* in which the agent's belief in G goes to 1 in each case is different. In the good case, it appears as if $\mathrm{cr}(G)$ goes to 1 as a result of a reliable belief-forming mechanism. In the bad case, $\mathrm{cr}(G)$ goes to 1 as a result of some form of wishful thinking. An account of evidence that is sensitive to the *way* in which a belief was produced, could make the appropriate distinction here. A similar thing holds for Case 2. Tom's full confidence that it is hot in New York is reliably produced, whereas Pete's full confidence that it is hot in Sydney is not reliably produced (given that Pete isn't in Sydney, and, as mentioned, that he has normal human faculties). An externalist account of what evidence is can handle both these cases.

Consider a different case, which tells in favor of some externalist account of evidence. In this case there are two agents, Alice and Beatrice, who are stipulated to be

---

resilient credence in P with value n. Joyce's suggested account would say that this is evidence for the agent when it is not.

internally identical. One can understand 'internal' in any way one likes, but for this example I will assume that they share all the same non-factive mental states. Now, imagine that Alice updates her belief state in an intuitively satisfying way. She receives evidence via her sensory organs, and then updates perfectly according to *COND*. Since Beatrice is Alice's internal twin, Beatrice also updates by conditionalization on the same propositions. However, whereas Alice believes her evidence propositions because she is using her sense organs to investigate the world around her, Beatrice is in a dark room and believes these same evidence propositions as a result of having her brain tinkered with by a maniacal scientist. Coincidentally, the scientist did things that put Beatrice into the exact same internal state as Alice. In this case I maintain that it is natural to say that Alice has different (specifically, more) evidence than Beatrice. However, by construction, this cannot be explained by appealing to internal features of Alice or Beatrice, since they are identical in this regard. So this third case not only *can* be handled by an external account of evidence, it is the *only* kind of account that has the resources to handle it. Altogether, this gives some reason to pursue an external account of evidence.

Of course, such preliminary considerations in favor of externalism about evidence are far from conclusive. In particular, one might argue that there are internalist accounts evidence that can handle Cases 1 and 2. A natural suggestion is an account according to which one's experiences determine one's evidence. This would not deliver what I claimed was the right response to the third case just given, but one could deny my reading of the case. It is a contentious matter, after all, what the correct response to the Alice and Beatrice case is, and one might claim that Alice and Beatrice in fact have

the same evidence. Considerations of internal duplicates have often been used to lead to the conclusion that various forms of externalism must be rejected.[8] So these cases in no way conclusively establish externalism about evidence.

In fact, I suspect that there are considerations telling in favor of both sides. When one thinks of evidence as the kind of thing that one always knows that one has, then internalism about evidence seems better motivated. For, the thought goes, if evidence is externally specified then I could have some evidence even though an internal twin of mine lacks it. But if that's the case then I know of some of my evidence that it is evidence. Another consideration for internalism concerns the notion of guidance. If we want to give an agent some rules of guidance to direct his epistemic behavior, then it would seem that his evidence better be internalist evidence, since internalist evidence seems to be the kind of thing he can use to guide his behavior.

On the other side there are considerations that tell in favor of externalist understandings of evidence. There is, it seems, a pretty strong intuition that evidence is true. We often speak of misleading evidence, but rarely of *false* evidence. For instance, Sherrilyn Roush argues ([2006], p. 173) that one important feature that evidence has is a high likelihood of being *true*. Others, such as Timothy Williamson ([2000]) insist that evidence must actually *be* true. One can understand why this is an important feature of evidence: we aren't going to pay much attention to alleged evidence for global warming or the alleged evidence for a suspect's guilt if it turns out that the evidence is false, or that it is very unlikely to be true.

[8] See, for instance, Silins ([2005]). I consider the relevant argument in this paper in Chapter 4.

Second, our evidence often is about things outside of our own heads. When we talk about the evidence for global warming this evidence does not simply describe features of an agent's internal state. Instead, the evidence is about features of the world. Consider evidence in a legal setting. The evidence for the defendant's guilt is not solely about each jury member's internal states. Rather, the evidence is about features of the world. Evidence, then, is often about the world, and not just a subjective internal state.

But these two ideas—that evidence is true and that evidence is about the world—imply that evidence is external. For imagine that I have as evidence that it is sunny outside. The secondd idea, that evidence is sometimes *about* the external world, supports the claim that this is often the case. But there is certainly an internal twin of mine in a situation where it is not sunny. Given the first idea about evidence, that evidence is very often true, this internal twin doesn't have that it is sunny as evidence. So, we're left with externalism about evidence.

Further, externalism about evidence can capture another intuition that we have about evidence: that one's evidence can be effected by bad environmental conditions. This is the intuition that as things get worse externally, things get worse evidentially.

All of this shows, I think, that externalism about evidence isn't simply a nonstarter. At least given what has been said so far, externalism and internalism about evidence seem to be on par, each answering to different intuitions. However, one might think, there are theoretical reasons to prefer an internalist account of evidence within the context of Bayesian Epistemology. In what follows, I will consider this suggestion. I will argue that Bayesian Epistemology, somewhat surprisingly, already embraces

externalist principles, and so we can pursue an externalist account of evidence unimpeded.

2.4 Bayesian Epistemology and Access Internalism

In order to profitably discuss internalism and externalism in epistemology, we must be careful to be clear about the theses. In what follows, I will introduce different versions of internalism, and take the negation of such views to be versions of externalism. The claim will be that orthodox Bayesian principles do not come out as internalist norms on two plausible ways of understanding internalism, and so should be understood as presenting us with externalist-friendly norms.

Internalism and externalism are often thought of as epistemic supervenience theses underwritten by some notion of access.[9] For instance, traditional internalism about justification holds, roughly, that S being justified in believing *P* supervenes on internal properties of S. But this is rough: we must restrict what sorts of internal properties are relevant. This is where the notion of access comes in. Many traditional internalists hold that the internal properties that are relevant are those that are introspectively accessible to S. Thus: S being justified supervenes on properties introspectively accessible to S.

This common way of understanding internalism and externalism doesn't work as nicely when we are considering epistemic rules or requirements, which is what we will be doing when we consider the principles of Bayesian Epistemology. It doesn't immediately make sense to ask whether or not an epistemic rule supervenes on internal

[9] See Pryor ([2001]). Pryor notes that internal facts are those to which the agent has a special kind of access.

properties of some epistemic agent. We can, of course, understand the *justification* of an epistemic rule according to the supervenience notion above, but this is not always what we are interested in when we are interested in whether a purported epistemic rule is internal or external.

We can, however, mold epistemic requirements into this framework. Consider *PROB*, the rule that says that, at any time, an agent should have a probabilistically coherent set of beliefs. Say that a belief set is *synchronically coherent* when it conforms to *PROB*. Then, we could propose the following:

> *PROB* is an internalist epistemic requirement iff the property of S being synchronically coherent at *t* supervenes on properties internally accessible to S at *t*.

This seems to capture well at least *one* sense of internalism, one that has to do with access. The Access Internalist says that all epistemic requirements must be internalist requirements:

> **Access Internalism:** Any epistemic requirement **R** must be such that the property of S being in conformance with **R** at *t* supervenes on properties internally accessible to S at *t*.

According to this, a requirement instructing S to believe the truth is not an an epistemic requirement. S believing the truth at *t* does not supervene on properties internally accessible to S at *t*. On the other hand, the requirement instructing S to believe what *seems* to be true is possibly an epistemic requirement. S believing what it seems to S to be true *does* supervene on properties internally accessible to S at *t* (so long as seemings like these are internally accessible).

The view can be further refined by specifying a certain sort of accessibility. A traditional way of doing this relies on the notion of introspection:

> **Introspective Access (IA) Internalism:** Any epistemic requirement **R** must be such that the property of S being in conformance with **R** at *t* supervenes on properties *introspectively* accessible to S at *t*.

A slightly different version of Access Internalism does not rely on introspective accessibility, but instead on cognitive accessibility:

> **Cognitive Access (CA) Internalism:** Any epistemic requirement **R** must be such that the property of S being in conformance with **R** at *t* supervenes on properties *cognitively accessible* to S at *t*.

CA-Internalism is different from IA-Internalism in that the properties relevant to epistemic requirements need not be introspectible, but rather only cognitively accessible. It is very plausible that there are things cognitively accessible to us, which we nevertheless cannot introspect. For example, I can't introspect certain properties of my visual experience, but there is reason to think that such properties are nevertheless accessible to various cognitive mechanisms that I have.

I will now consider three examples, which show that Bayesian Epistemology is not committed to either IA-Internalism or CA-Internalism.

2.4.1 Case 3

Consider Bob. At some time, *t*, Bob randomly happens upon a coherent credence function, although at all other times, his beliefs are not coherent. Imagine that Grant has a credence function that is identical to Bob's at *t*, but that Grant reached that epistemic state via conditionalization. At *t*, Bob and Grant have epistemic states with

the same synchronic properties. But, if one thinks that *COND* is an epistemic requirement, then one thinks that there is an epistemically significant difference between Bob and Grant. And indeed, it is plausible that there is such a difference.

COND identifies what this difference is: Grant's epistemic state has the property, *resulting-from-conditionalization*, whereas Bob's epistemic state has the property, *not-resulting-from-conditionalization*. This property, which makes Bob epistemically worse off than Grant, is not something to which Bob has any sort of internal access. And similarly, Grant doesn't have access to the property which makes him epistemically better off than Bob. Bob and Grant don't have access to these properties because *COND* concerns a relation between one's *past* epistemic state and one's *present* epistemic state. But Bob's past epistemic state is not internally accessible to Bob (and the same for Grant). At best, Bob has internal access to his *present* epistemic state. This present epistemic state may include *memories* of his past epistemic state, but this is not to say that it *includes* his past epistemic state. Thus, *COND* is not IA- or CA-Internalist. For Bob does not have access to the properties that make him fail to fulfill *COND*. If we really wanted an internally accessible requirement, then we'd have to say that what matters is not that Bob's *actual* past epistemic state is related by *COND* to Bob's current epistemic state, but rather that Bob's current *memories* of his past epistemic states are related by *COND* to Bob's current epistemic state.[10] This, however, is not the kind of constraint laid down by *COND*. So, acceptance of *COND* as

[10] Chris Meacham also makes this point in his ([*forthcoming*]).

an epistemic requirement already commits one to an external epistemic requirement, which appeals to things that may not be internally accessible to the agent in question.[11]

2.4.2 Case 4

The second example concerns *PROB*, and shows that internal access has nothing to do with the Bayesian norms. Imagine that Kate makes the following objection to the Bayesian requirement that her beliefs be coherent (*PROB*):

> "But I don't know what all my degrees of belief *are*, so I have no way of making sure that they are coherent."

A standard response to Kate's complaint is that it is simply irrelevant to the claims that Bayesian Epistemology is making.[12] But why is this irrelevant?[13] In Chapter 1, we saw

---

[11] Note that this example also shows that *COND* is not a species of a non-access version of internalism. Consider a form of internalism says that a requirement is internal iff it concerns one's present mental state. Since *COND* doesn't concern one's present mental state, it is not an internalist requirement. I do not explicitly discuss this, since it seems to me that this restriction to one's present mental state is itself motivated by considerations of access. In particular, by the idea that one has access to all and only those things in one's present mental state.

[12] Mark Kaplan gives a response like this in his ([1996]). He considers the fact that a view he calls Modest Probabilism, which includes *PROB* as a part, requires one to assign tautologies maximal credence. He then imagines an objection: "…[Modest Probabilism] would require you to survey the set of all hypotheses so as to pick out the tautologies. But what counts for you as the set of all hypotheses is presumably sufficiently large that you could not possibly succeed in surveying all the hypotheses it contains." ([1996], p. 36) His response is telling:

> Where Modest Probabilism properly understood as a rule for the conduct of inquiry which you must satisfy in its every detail on pain of being called on the carpet, then the problems would be very serious indeed. But were that so, it would be odd to speak of Modest Probabilism as a regulative ideal…Being a regulative ideal, Modest Probabilism is not to be understood as giving you a set of marching orders which you can violate only at the cost of being classified as irrational. Modest Probabilism, is, rather, to be understood as putting forth a standard by which to judge the cogency of a state of opinion… ([1996], p. 37)

Although David Christensen never addresses this particular issues, he appears to understand *PROB* in a way that is consistent with this response. He gives an example of Kelly, who has an incoherent set of beliefs and Mark who has a coherent set. He then writes: "Mark is (all else being equal) more rational than Kelly: his degrees of belief fit together in a way that respects the logical interconnections among the claims believed. And this is so even if, owing to her psychological makeup, Kelly is incapable of doing better cognitively." ([2004], p. 177) Kate may be psychologically incapable of knowing all her credences, but this does not in any way impugn on *PROB* as a rational requirement.

that Bayesian Epistemology can be understood as giving us theories designed to evaluate the epistemic functioning of an agent. From the fact that Kate lacks access to all her own degrees of belief, nothing follows about whether or not it is an epistemically good thing that they be coherent. But that is to say that *PROB* is indifferent to whether or not an agent has internal access to the things that *PROB* requires. But this is to see *PROB* as an externalist requirement, one that is binding on Kate even if she lacks access to the things on which her fulfillment of *PROB* supervenes.

2.4.3 Case 5

The final example comes from reflecting on the larger philosophical project of which Bayesian Epistemology makes up the epistemic part. A Bayesian account of rational belief is naturally paired with a decision theoretic account of rational decision-making. Sometimes this is necessarily the case if the notion of degree of belief or credence is defined in terms of rational betting preferences. But even if the two are not as tightly related as this, there is still a close relationship. This is, for instance, why Dutch Book Arguments have been thought to tell in favor of some of the Bayesian constraints on rational belief.[14] But the decision-theoretic account of rational decision-making lays down requirements on rational decision-making that are not access internalist requirements.

---

[13] One way of answering this question is to see *PROB* not as a normative requirement, but rather as a precondition for Bayesian Epistemology to apply to an agent. However, as noted in the introduction, I'm not understanding *PROB* in this way. Thus, this question needs to be answered in a different way.

[14] Dutch Book Arguments go back at least to F. P. Ramsey's classic paper, "Truth and Probability" ([1926/1990]). The argument can be found other places, including de Finetti ([1937/1980]), Teller ([1973]), Skyrms ([1975]), Horwich ([1982]), Skyrms ([1987]), Sobel ([1990]), Armendt ([1992]), Skyrms ([1993]), Howson & Urbach ([1993]), Lewis ([1999]), McGee ([1999]) and Briggs ([2009]). Criticism of Dutch Book Arguments can be found in Bacchus, Kyberg, & Thylos ([1988]), Christensen

According to the basics of that account, the rational decision is the one that maximizes subjective expected utility. However, an agent can make the rational decision (choose the option that maximizes subjective expected utility) for the completely wrong reasons. For example, say that Jim chooses one of three stock portfolios purely because he always chooses the first thing presented to him. Nevertheless, if the first of the three stock portfolios maximizes Jim's subjective expected utility, then the Bayesian account of rational decision making says that Jim made a rational decision. This is true even if Jim lacks access to some of his credences and utilities. Thus, we have an external requirement on rational decision-making that does not appeal to things accessible to Jim to justify his decision, but rather appeals to general features of Jim's situation that made Jim's decision the best one.

One might object that these examples show what I claim they do. With respect to examples 5 and 6, for instance, one might claim that features of an agent's credence function and utility function at a time *are* in fact accessible to that agent at that time. Alternatively, one might claim that, to the extent that features are *not* accessible to that agent at that time, then the norms no longer apply. This is a hard line to take, however. For consider the property that an agent's total credence function has when it satisfies *PROB*. Depending on the computational abilities of the agent, this feature may not be accessible to the agent. This rules out the first response. With respect to the second response, we would have to say that *PROB* no longer applies to that agent. But this is just what we wanted to avoid by understanding Bayesian Epistemology as giving evaluative, anti-procedural theories.

([1991]), Maher ([1992]), Howson ([1997]), Levi ([2002]), and Williamson, J. ([*forthcoming*]). For more on Dutch Book Arguments, see Chapter 6.

Even if this response could be maintained, however, we would still have the problem with *COND*. Since part of what makes for satisfaction of *COND* (the agent's past doxastic state) is no longer around, such a feature is not accessible. Thus, such a norm is not Access Internalist.

Summing up this section, I think we have good reason to think that the Bayesian orthodoxy does not endorse any kind of Access Internalism. Whether or not an agent has access to the property of being in conformance with a Bayesian principle, the principle is still binding. So, we can conclude that Bayesian Epistemology is an Access *Externalist* theory. This is important. For if one's reason for resisting an external account of evidence turned on considerations of access—for instance, the idea that such evidence may not be internally accessible, or that one may not know what one's evidence is—then this shows that such a motivation mandates a broad rejection of key Bayesian principles.

2.5 Guidance Internalism

In response to the arguments in the previous section, one might claim that there is a different kind of internalism, still reliant on the notion of access, which will show the Bayesian requirements to be internalist. Above, it was important that the agent had access to whether or not she had met the relevant requirement.

However, one might claim that this is not why access is important. Instead, access is important to ensure that the epistemic requirements are capable of *guiding* the agents to which the requirements apply. The idea is that even if the property of coherence that an agent's total doxastic state has (or lacks) at a time is not accessible to the agent, if the credence value for each proposition is accessible to the agent, then this

is enough to ensure that something like *PROB* could guide the agent. Similarly, even if an agent's past doxastic state is not now accessible to the agent, it is enough that the past doxastic state *was* accessible in a way to allow *COND* to guide the agent. It is this notion of guidance that is important, the thought goes, not access to whether one conforms to the principles.

We can attempt to formulate this kind of internalism as follows:

> **Guidance Internalism:** Any epistemic requirement **R** must be such that it is possible that **R** guides agent S.

Clearly, Guidance Internalism is not very informative until we have said something about what it is for it to be possible that **R** guides S. I think the following gives at least a necessary condition:

> If it is possible that requirement **R** guides agent S, then there is a set of properties P such that (i) S responding to the members of P guarantees **R**-conformance and (ii) the members of P supervene on properties internally accessible to S.

Now, as is clear from the definitions above, Guidance Internalism still rests on some notion of access. However, the reason for the required access is different. Above, we wanted access so that the agent was in a position to tell if he was satisfying the requirements. Here we want access to ensure that the agent is in a position to be guided by the requirement.

Guidance Internalism is a popular position. John Pollock's discussion of internalism and externalism, for instance, seems to bring out the importance of something like Guidance Internalism. In his ([2000]), he distinguishes two sorts of

externalism to be contrasted with the internalist approach to epistemology that he favors. Pollock contrasts what he calls *belief externalism* with *norm externalism*:

> *Belief externalism* insists that correct epistemic norms must be formulated in terms of external considerations. A typical example of such a proposed norm might be "It is permissible to hold a belief if it is generated by a reliable cognitive process." In contrast to this, *norm externalism* acknowledges that the content of our epistemic norms must be internalist, but employs external considerations in the selection of the norms themselves. ([2000], p.193)

Pollock rejects both of these forms of externalism, but he seems to think that norm externalism is more defensible that belief externalism. At any rate, his discussion highlights the fact that one can deny belief externalism while still holding on to norm externalism. Why, then, might one think that norm externalism is preferable to belief externalism, or why distinguish them at all? The reason, I think, comes down to the notion of guidance. A Norm Externalist requirement, though external in some sense, is still the kind of requirement that could *guide* an epistemic agent. If following a certain requirement only requires the agent to respond to a set of properties, and if this set of properties is internally accessible to the agent, then an epistemic agent can be guided by that rule so long as the agent monitors and responds to those internally accessible properties. If all epistemic requirements were like this, then good epistemic behavior would be like following an algorithm. We could, in principle, build perfect epistemic robots.[15]

An internalism much like Guidance Internalism is explored at length by Ralph Wedgwood in his ([2002]). In that paper, Wedgwood talks about *basic rules*. Basic rules are rules that we follow "directly", without following any other rules. In this sense, they are analogous to basic actions. Wedgwood writes: "If a thinker is able to

follow a certain rule directly, in this sense, at a given time, then I shall say that the rule in question is a "basic rule" for that thinker at that time." (p. 355) But Wedgwood holds that the only rules we can follow directly are rules that appeal to internal features of our mental lives.[16] So, the only basic rules are internal rules. Since Wedgwood holds that rules of rationality are basic rules[17], such rules are internal.

The motivation behind this kind of internalism is subtly different from the motivation behind Access Internalism. The motivation for Access Internalism turns on the idea that epistemic agents must be able to tell whether they're meeting their epistemic requirements. On the other hand, the motivation for Guidance Internalism is the idea that epistemic requirements must be capable of guiding our epistemic moves. This motivation emerges from Wedgwood's discussion when he describes what it is to conform to an epistemic rule in the sense that he is concerned with:

> As I have described these rules, it is perfectly possible to conform to these rules by pure fluke. But if it is purely a fluke that one conforms to a rule, it will hardly be appropriate to say, even metaphorically, that conforming to the rules is something that one does, in order to pursue this aim. This description will be appropriate only if one not only conforms to the rule, but also **follows**, or is **guided by**, the rule. (p. 354)

---

[15] Indeed, this general outlook fits in well with Pollack's OSCAR project (Pollock, [1995]).

[16] Wedgwood's defense of this is quite complex. Here's the basic idea:

> Suppose that one directly follows a basic rule that permits one to revise one's beliefs in a certain way, whenever one is in a certain condition. Then, as I explained in §3, a correct fully-articulated explanation will identify the fact that one is in that condition as at least part of the *proximate explanation* of that belief revision. So, the fact that one is in this condition must itself be an "internal fact" about one's mental states. In general, following such basic rules always involves revising one's beliefs in response to such internal facts. This is not to say that it is impossible to follow a rule that permits one to revise one's beliefs in a certain way whenever a certain *external* fact obtains. It may be quite possible, for example, to follow the rule that permits one to believe *p* whenever one can *see* that *p* is the case. But this rule cannot be a "basic rule". If one follows this rule, one does so *by means of* following some basic rule, such as the rule that permits one to believe *p* whenever one has a visual experience as of *p*'s being the case (and no reason to distrust one's experience in the circumstances). (p. 364)

[17] "If the thinker revises her belief in *p* at that time, then that belief revision is rational just in case it results from her directly following some of these basic rules that it "makes sense" for her to conform to." (p. 354)

A similar sentiment can be seen in John Pollock's discussion of epistemic norms. He writes:

> The internalization of norms results in our having "automatic" procedural knowledge that enables us to do something without having to think about how to do it. It is this process that I am calling "being guided by the norm without having to think about the norm." This may be a slightly misleading way of talking, because it suggest that somewhere in our heads there is a mental representation of the norm and that mental representation is doing the guiding. Perhaps it would be less misleading to say that our behavior is being guided by our procedural knowledge and the way in which it is being guided is described by the norm. What is important is that this is a particular way of being **guided**. It involves nonintellectual psychological mechanisms that both **guide** and **correct** (or fine tune) our behavior. ([1999], p. 195-6, my boldface)

Both Pollock and Wedgwood make clear that underlying their internalism is the idea that the epistemically important things—whether norms or requirements (or anything else)—are the sorts of things that *guide* epistemic behavior.

2.5.1 Guidance Internalism and *COND*

Though Guidance Internalism is subtly different from the Access Internalism described in Section 2.4, it is enough to make a difference to our discussion. Consider *COND*. In Section 2.4, Access Internalism required that the epistemic agent have access to whether he had met the requirements of *COND*. *COND* clearly failed this, since at *t* you don't have access to whether you have met the requirements of *COND*. Nevertheless, it is plausible that an agent could be internally guided by a *COND*-like algorithm. Although it is not internally accessible at *t* whether one has conformed to *COND*, there seem to be "basic rules" that one can follow (or, nonintellectual psychological mechanisms one could have), which only require one to respond at *t* to properties that are internally accessible at *t*, and that guarantee conformance to *COND*.

To see this how this might go, assume for the moment that S's evidence at *t* is internally accessible and that the property of being S's evidence at *t* supervenes on what is internally accessible at *t*. At *t*, S has a credence function, cr_t. S then receives some evidence, *E*. Since *E* motivates the change in S's credence function, before S receives *E*, S's credence function is cr_t. Upon receiving evidence *E*, there is a moment at which S is in possession of evidence *E*, and S's credence function is still cr_t. Thus, S has access to both *E* and cr_t. The evidence and the credence function then interact appropriately to yield S's new credence function. Thus, it seems to be possible for S to follow *COND* by, at every moment, correctly responding to S's credence function and evidence at that moment, both of which are accessible to S at that moment. That is, *COND* gives a kind of algorithm that can be internally followed—whether at a personal level as a basic rule, or at a subpersonal level by way of a psychological mechanism. Thus, *COND* seems to be the kind of thing that could internally *guide* an epistemic agent. For this kind of internalism, I propose, *COND* looks to be an internalist requirement.[18]

One could object to this. In arguing that *COND* could guide S, I claimed that there is some moment at which S has evidence *E* and yet S's credence function is still cr_t. It's not immediately obvious how this is to be understood. This is easiest to see if we assume that a necessary condition on having evidence *E* is that $cr(E) = 1$. If $cr_t(E) \neq 1$, then it is not possible for S to both have *E* and yet $cr_t(E) = 1$. If $cr_t(E) = 1$, then if the agent is coherent, receiving evidence *E* has no effect. One might claim, then, that there is no moment at which S is in possession of *E* and S's credence function is still cr_t.

[18] A similar thing holds for *PROB*. Although the agent may not be able to tell that she has a coherent credence function, she may have access to all the features of her credence function that go into making it coherent. Thus *PROB* is the sort of rule that could internally *guide* an epistemic agent, though this would most likely go on at a sub-personal level.

Instead, the receipt of evidence and the update are simultaneous. Given this, there is no moment at which the agent has both the ingredients that go into the new, updated credence function. Instead, the agent's credence function simply evolves from moment to moment. This is an interesting objection to seeing *COND* as a Guidance Internalist requirement. Perhaps it is successful. If so, then *COND* shows that the Bayesian norms are not Guidance Internalist norms, just as it shows that they are not Access Internalist norms. But perhaps this objection is not successful. If not, then we will need a further reason why the Bayesian norms need not be seen as Guidance Internalist norms. In the next section I will provide such a reason.

2.5.2 Guidance Internalism and Evidence

It seems that there might be a way of understanding internalism—Guidance Internalism—according to which the traditional Bayesian requirements come out as internalist requirements. Above I said that the orthodox Bayesian should reject Access Internalism since key Bayesian requirements seemed to violate it. But the same kind of response to Guidance Internalism may not be available, and commitment to Guidance Internalism can be used to argue against an external account of having evidence. If what evidence *is* is partly an external matter, then S need not have access to what S's evidence is. That is, the property of being S's evidence doesn't supervene on properties accessible to S. But if *COND* is to guide S, this requires S to have cognitive access to the property of being S's evidence. If we assume a commitment to Guidance Internalism, what S's evidence is must be cognitively accessible, and thus not externally specified. So, it seems as though a commitment to Guidance Internalism tells against an external account of evidence. For, the thought goes, if what evidence *is* is not something

that I am in a position to determine, then how can evidence guide me in my epistemic endeavors?

I think that this line of argument is evident in several prominent writers. Consider, for instance, Richard Jeffrey's ([1965]) classic *The Logic of Decision*. In Chapter 11, when Jeffrey motivates his generalization of conditionalization (so-called "Jeffrey Conditionalization") he appeals to examples about observation under candlelight. He argues that in such a situation, there is no evidence proposition *E*, which adequately describes the agent's experience, such that an agent can conditionalize on *E*. He writes:

> But there need be no such proposition *E* in his preference ranking; nor need any such proposition be expressible in the English language. Thus, the description "The cloth looked green or possibly blue or conceivably violet," would be too vague to convey the precise quality of the experience. ([1965], p. 165)

His response to this problem is his famous generalization of conditionalization. Underlying this whole discussion, however, is the idea that, whatever evidence is, it is something about the agent's experiences, that the agent has access to, and indeed can express to himself. This position makes perfect sense for Jeffrey, as he makes clear a few pages later: "To serve its normative function, the theory of decision making must be **used** by the agent, who must therefore be able to formulate and understand the relevant propositions." ([1965], p. 167, my boldface)

Consider also Paul Horwich's ([1982]) position in his book, *Probability and Evidence*. Horwich claims he doesn't go for a positivistic, foundationalist view about evidence, but thinks that evidence is rather contextually determined (pp. 38-42). Nevertheless, he also says that rational constraints must be internal constraints:

> What I mean by an internal constraint is one whose application depends only upon the intrinsic content of an epistemic state. Consistency is an internal constraint; for in order to determine whether or not it is satisfied, it suffices to know simply which beliefs are contained in the epistemic state. One need not know, for example, how the beliefs were caused, or anything about the previous history of the subject. An external constraint, on the other hand, implies that the rationality of a system of beliefs depends on circumstances which go beyond the identity of the beliefs. It has been suggested, for example, that the rationality of certain kinds of belief may depend upon their generation by a reliable belief-producing mechanism – a process, such as visual perception, which tends to result in true beliefs – regardless of the subject's awareness of this fact. ([1982], pp. 75-6)

Horwich then proceeds to argue that rationality concerns *internal* constraints, rather than external constraints:

> …rationality is a normative concept; the adoption of beliefs in the absence of justifiability is wrong and subject to disapprobation. But such an attitude would be quite inappropriate if the conditions for justification included circumstances whose presence we could not be expected to recognize… Insofar as the rationality of a system of beliefs depends simply upon its having properties whose possession may be inferred from the identity of the beliefs, then we can reasonably expect the subject, who is aware of his beliefs, to determine whether or not they are rational. But if extrinsic constraints govern justification…then there arises the possibility of a reasonable, yet incorrect, estimate of the presence of those conditions… ([1982], pp. 76-77)

Note that Horwich begins giving a sort of deontological argument for internalism, citing the wrongness of believing without justification. Deontological arguments such as this often go hand-in-hand with commitment to Guidance Internalism. For, the thought goes, if there's nothing to *guide* us towards meeting a requirement, then there can't be anything wrong with failing to meet it. But since there *is* something wrong with failing the requirements (the deontological idea), there must be something to *guide* us to meet our epistemic requirements.

Outside of Bayesian circles, this idea is no less prominent, as Timothy Williamson has noted. Williamson refers to the basic idea behind Guidance Internalism as 'Operational Epistemology'. He writes:

> Operational epistemologists may allow some value to non-operational epistemology, in third-personal assessments of enquiry which take into account facts unavailable to the enquirers under assessment. But, they argue, we should also do epistemology in a first-personal way, in which epistemologists think of the enquirers as themselves. More precisely, it should provide enquirers with **guidance** which they can actually use, in whatever situation they find themselves. (Williamson, [*forthcoming*])

Guidance Internalism, then, appears to be a popular way of understanding internalism. And at least potentially, it provides a way to understand Bayesian principles according to which they come out as *internalist* principles.

2.5.3 Response

Though Guidance Internalism is a popular sort of position, and though commitment to it instructs us against giving an externalist account of evidence, it cannot motivate the Bayesian Epistemologist who accepts Bayesian Epistemology as an evaluative, anti-procedural theory to reject an externalist account of evidence. The reason for this is simple. The evaluative, anti-procedural understanding of Bayesian Epistemology says that we are not attempting to give a manual for epistemic performance, nor rules that the agent must be capable of following. Guidance Internalism, however, says that the epistemic rules and requirements that we lay down should be capable of guiding agents in their epistemic performance. You cannot accept both of these. So, if we hold onto the understanding of Bayesian Epistemology from Chapter 1, then we must reject Guidance Internalism, and with it the reason to reject an externalist account of evidence.

Given this, we are to understand *PROB* and *COND* as laying down evaluative standards, which need not provide guidance or instruction to the agents in question. What do such principles do for us? They allow us to evaluate how well some epistemic agent did in an objective sense, even if the agent doesn't do well *by* following some

PROB or *COND* algorithm.[19] If we understand Bayesian Epistemology in this way, however, then we are free to investigate an account of evidence that fits the same mold: an account of evidence that need not provide guidance or instruction, but rather serves as a standard of evaluation.

In fact, Guidance Internalism must be rejected if we are to pursue an evaluative, anti-procedural Bayesian Epistemology that can get around the OIC argument. For imagine that it *is* a condition of adequacy on epistemic principles that they can guide the epistemic agents towards whom the theory is directed. There would then be a quick argument, very similar to the OIC argument, telling against *PROB* as an epistemic requirement:

(1) If *PROB* is an epistemic requirement, then it is possible that *PROB* guide human agents.

(2) It is not possible that *PROB* guide human agents.

(C) Thus, *PROB* is not an epistemic requirement.

Since premise (2) is very plausible for the computational reasons sketched in Chapter 1, the defender of *PROB* as an epistemic requirement must deny premise (1). But to reject premise (1) is to claim that epistemic requirements need not be such that they can guide epistemic agents. That is, to deny premise (1) is to say that *PROB* can be an epistemic requirement even if it cannot guide human agents.

---

[19] Indeed, David Christensen ([2004]) writes:

> Alston argues that what we really are interested in when we theorize about epistemic justification is a non-deontological but still clearly *evaluative* notion…And it seems to me that this sort of understanding is even more plausible when applied to epistemic rationality. (p. 161)

And he continues this thought a page later:

> Rationality is a good thing, like sanity, or intelligence. It is not the same thing as sanity, or as intelligence, but it is more similar to these notions than it is to notions involving obligation or deserved blame. We may call a person's beliefs irrational without implying that she had the capacity to avoid them…In epistemology, as in various other arenas, we need not grade on effort. (p. 162)

2.5.4 Objection

One might object to this and say that even though principles like *PROB* and *COND* (when paired with an internal account of evidence) are not intended to guide human agents in their epistemic endeavors they are still such that they *could* be used by certain computationally advanced agents for guidance. When we pair *COND* with an external account of evidence, on the other hand, it cannot be used by *any* kind of agent for guidance. For if there were an agent that could use externally specified evidence for guidance, then the agent would have the appropriate access to the evidence, and so for that agent the evidence would not be externally specified. Thus, the thought goes, an external account of evidence is fundamentally different than something like *PROB* or *COND* (when it is paired with an internal account of evidence). The basic idea here is that if *COND* is specified with internal evidence, then there is *some* agent such that *COND* can guide that agent. Perhaps the agent is very much unlike you or I in powers of introspection and computation, but nevertheless, there is such an agent. However, the thought goes, if *COND* is specified with external evidence, then there is not some agent such that *COND* can guide that agent. This is because if the evidence is specified in an external way, then by definition such evidence cannot guide that agent.

There are two problems with this line of response. First, there is no reason why it would matter to the acceptability of an epistemic requirement that it is capable, in principle, of guiding some agent *different* than the agents we're interested in. This response says that it makes a difference to an epistemic requirement on human agents that there is some agent that could be guided by the requirement. But why would that matter?

Second, there is an important ambiguity in the claim that if *COND* is specified with external evidence, then there is no agent such that *COND* can guide that agent. In particular, this claim can be understood in two ways:

(a) If *COND* is paired with an account of evidence where the evidence is external with respect to agent A, then *COND* cannot guide agent A.

(b) If *COND* is paired with an account of evidence where the evidence is external with respect to agent A, then there is no agent such that *COND* can guide that agent.

(a) is true, but (b) is not. Evidence that is specified externally with respect to A may not be external with respect to some other agent with different cognitive abilities. For instance, that the pan is hot may be internally accessible to some agent, even if it is not internally accessible to agent A, and so is external with respect to A. But if (a) is what is being claimed, then the same line of thought tells against a requirement like *PROB*. For consider:

(c) If *PROB* is computationally intractable for agent A, then *PROB* cannot guide agent A.

(c) is true for the same reason that the conditional in (a) is true. So, if the truth of (a) tells against the viability of an external account of evidence, then the truth of (c) tells against the viability of *PROB* as an epistemic requirement. But we are understanding Bayesian Epistemology in such a way that the truth of (c) does not pose a problem. So the truth of (a) must not pose a problem either.

The upshot of all of this is that *if* one sees Bayesian Epistemology as a kind of evaluative, anti-procedural theory, then one has already rejected Guidance Internalism.

If one does this, however, then one loses the motivation for a Guidance Internalist account of evidence, opening the door to an externalist account. This is important. If one's reason for rejecting an externalist account of evidence turned on Guidance Internalist considerations, then this is no reason to reject such an account when it comes to Bayesian Epistemology, as it is being understood here.

2.6 A Different Understanding of Guidance Internalism

I have attempted to show that Bayesian Epistemology, as understood in Chapter 1, is not committed to internalist principles. I first showed this for Access Internalism, and have now shown this for Guidance Internalism. However, there is a different way of understanding internalism, which still works with the notion of guidance, according to which we might reach a different verdict.

An epistemic requirement or norm is Guidance Internalist just in case it is possible that the norm or requirement guide the agent in question. Accordingly, a commitment to Guidance Internalism says that the epistemic norms our theories appeal to must be capable of guiding the epistemic agents to which they apply. An alternative formulation of Guidance Internalism, however, says something rather different:

> **Guidance Internalism***: Any epistemic requirement **R** must be such that when **R** is applied, any two internally identical epistemic systems are evaluated identically.[20]

[20] Thanks to David Christensen for suggesting this alternative way of understanding Guidance Internalism.

The idea here is that we want our epistemic requirements to give an indication of how well the agent's epistemic system is doing, while ignoring factors outside of the agent (for instance, factors in his environment).

GI* is a recognizably internalist constraint, but one might wonder what it has to do with guidance. To see this, imagine that somewhere within their cognitive equipment, agents have an epistemic module. This epistemic module takes care of all the belief-formation and belief-processing for the agent. A sensible idea is that the epistemic module, just like the rest of the agent's mind, is computational. As a computational system, the epistemic module directly responds to only proximal stimuli. But proximal stimuli for such an epistemic module will all be internal features of the agent's cognitive system. In fact, most of the things directly affecting this epistemic module will be internal to the epistemic module—e.g., other beliefs. The idea behind GI* is that we should see the agent's epistemic module as the epistemic guidance system for the agent in question. Adherence to GI* guarantees that we're evaluating these epistemic modules—these epistemic guidance systems—based only on intrinsic features of the modules themselves.

Notice that traditional Bayesian requirements seem to satisfy GI*. This is easiest to see with *PROB*. Since identical epistemic systems will be identical with respect to their credence functions, *PROB* will deliver identical verdicts for agents with identical epistemic systems.[21] *COND* satisfies GI*, too, but only if 'identical epistemic systems' refers to epistemic systems that are identical at *all* times. If we understand things in this way, then *COND* is a requirement that satisfies GI*. Notice also that the

evaluative, anti-procedural understanding of Bayesian Epistemology is consistent with GI*, for it is consistent with this constraint that we ignore computational and other internal limitations, so long as we do this for *all* agents with identical epistemic systems.

If, however, we were to add an external account of evidence, then the resulting Bayesian requirements would not satisfy GI*. This is because there could be two agents with identical epistemic systems, but with different evidence on account of different environmental factors. Accordingly, they would be evaluated differently by the requirement that agents update on their evidence. Thus, it might appear that we have a problem for external accounts of evidence. The problem can be put in argument form as follows:

(1) Bayesian requirements must satisfy Guidance Internalism*.

(2) Any external account of evidence fails to satisfy Guidance Internalism*.

(C) Thus, any external account of evidence is not a Bayesian requirement.

I've just explained why Premise (2) is true. Premise (1) is supported by the fact that *PROB* and *COND*, the dominant Bayesian norms, meet GI*. In fact, other examples could be produced to provide more support for Premise (1). For example, a popular Bayesian principle is Lewis's Principal Principle:

PP: $\mathrm{cr}(P|\mathrm{p}(P) = n) = n$

where p(•) is an objective probability function. One can verify that such a requirement also satisfies GI*.

---

[21] If we broadened our view to encompass epistemic+motivational systems, a similar thing would be true of the requirement to maximize SEU. Since I am focusing purely on epistemic matters, I will not go through this explicitly.

What can be said in response to this argument? The response is that it is incorrect to see Bayesian Epistemology as committed to GI*, even though its standard requirements are consistent with it. That is, though the most popular Bayesian requirements *do* satisfy GI*, this tells us nothing about what the other Bayesian requirements must be like. This is because *any* epistemic theory will have to say something about the mental features of agents, in particular features of their epistemic systems. Standard Bayesian norms focus on this particular aspect of the theory, so it is no surprise that these norms treat identical epistemic modules the same: all the norms are concerned with is the inner workings of such modules! However, the fact that these norms are neither Access Internalist nor Guidance Internalist is reason to believe that this focus is incidental and not essential to Bayesian Epistemology. Further, when we come to the project of saying what it is to have evidence, this is exactly where we'd expect to appeal to features of the agent or his situation that are external to his epistemic module. This is because it is the receipt of evidence where the agent makes contact with the world. So, Premise 1 can be reasonably denied: we have a reason why the standard Bayesian requirements satisfy GI*, that doesn't lend support to GI* as a general methodological principle.

Note further that the standard Bayesian norms only concern features of the *belief state* of the agent in question. But if we are to give an account of evidence that goes beyond the deflationary picture, we must appeal to features of the agent external to his belief state. Thus, *any* non-deflationary account of evidence will require us to have norms that are sensitive to features *external* to the features that the standard Bayesian

norms are sensitive to. Once we appeal to such features, the Bayesian norms provide no support for a restriction about which features our norms can and cannot be sensitive to.

In spite of this, one might press on objecting to externalism about evidence on the basis of GI*. At this point, however, the charge against external accounts of evidence has changed significantly. Instead of claiming that Bayesian Epistemology is committed to GI*, the objector is claiming that something like GI* is simply a general constraint on epistemic theorizing. The discussion is thus a discussion of internalism and externalism in epistemology *in general*, and has nothing to do with Bayesian Epistemology *per se*. Of course if one thinks that externalism in general is a bad idea, then one will think that an externalist extension of Bayesian Epistemology is a bad idea. But, importantly, this is no reason from within Bayesian Epistemology to reject the addition of externalist requirements.

Nevertheless, is there reason to endorse GI* in general? One plausible reason in favor of GI* has to do with wanting to ignore cases of epistemic luck. That is, we might want to evaluate agents independent of contingent features of their environment that affected their belief-forming performance. So, for instance, one might want to evaluate two epistemic agents in the same way, even though one is looking at real barns and the other is looking at barn façades.[22]

I grant that it is sometimes useful and appropriate to evaluate epistemic systems in this way, abstracting away from features of the environment in which beliefs are being formed. However, it seems that pointing to situations of epistemic luck is a dangerous way to argue for GI*, if one is pursuing a non-deflationary account of

[22] David Christensen proposed this motivation in private correspondence.

evidence. When we have a case of external epistemic luck, like in the barn façade example, we have two agents that are mentally identical and yet one succeeds and the other fails in a certain task because of some feature of the environment external to the agents. The guiding idea is that such environmental differences should not lead us to epistemically differentiate the two agents. But consider a case where two agents are *doxastically* identical and yet one succeeds and the other fails at a certain task because of some non-doxastic feature of their mental state. For example, perhaps they both become confident that the light appears red, but it really only appears red to one of them. Nevertheless, they are both confident that it appears red to them. If we push on the idea that we want to eliminate cases of epistemic luck—cases where you get something wrong, but where this is outside your control—then we should eliminate this case of epistemic luck, too. The result of this is that we are pushed back towards a deflationary account of evidence. So, trying to motivate GI* based on the idea that we want to control for epistemic luck is a dangerous prospect if we hope for a non-deflationary account of evidence.

Let me be clear about the claim I am making. I am not saying that one cannot pursue a version of Bayesian Epistemology consistent with GI*. For when we're giving an evaluative theory, we are evaluating something, and it makes perfect sense to evaluate something while restricting attention to things internal to that thing. Obviously, we can evaluate an agent's epistemic performance while restricting attention to those things that are internal to the agent. Here's an analogy[23]: we might want to evaluate the performance of different cars, while ignoring particular details of their environments.

[23] This analogy was suggested to me by Chris Meacham.

So, we evaluate them in a purely internal way, say by looking at their horsepower. This displays the obvious point that evaluation does not entail *external* evaluation. So, the fact that we're giving an evaluative account doesn't automatically tell us that we must be giving an external account. But note, equally, that such an analogy does not show that external evaluation is incorrect, either. For, we can (and do) evaluate different cars in external ways, in terms of their longevity, or their top speeds, or their abilities to handle different weather conditions, all of which are dependent on external conditions. So, while evaluation doesn't entail external evaluation, it also doesn't tell against it.

2.7 Conclusion

I think we've reached an interesting conclusion: commitment to the standard principles of Bayesian Epistemology do require commitment to Access Internalism, Guidance Internalism, or Guidance Internalism*. Thus, one cannot point to the standard Bayesian principles as a reason to reject externalism about evidence. This, together with the *prima facie* reasons in favor of an externalist account of evidence (surveyed in Section 2.2), provides strong motivation for the Bayesian Epistemologist to investigate an externalist account of evidence.

But let me be honest: I have not argued that Bayesians *must* be committed to externalist accounts of evidence. Instead, I have suggested that externalism about evidence sits well with other Bayesian principles. But it is possible that the right response to the considerations I've raised here is to *change* the Bayesian principles to make them more internal, rather than continue with these externalist-friendly principles. Perhaps we should only pursue extensions of BE that are consistent with Guidance Internalism*.

Perhaps. But perhaps not. Perhaps there is something profitable to be gained by exploring the consequences that such an addition to Bayesian Epistemology might make. The rest of this dissertation will engage in this exploration. In the next chapter (Chapter 3) I will propose a specific externalist account of evidence. In Chapter 4 I will defend this account from a myriad of specific objections, some of which will lead to refinement of the view. In Chapter 5 I explain how the view I propose can be used to explain how agents could lose evidence. Finally, in Chapters 6 and 7 I apply my account of evidence to Dutch Book Arguments and to issues that arise with second-order doxastic states.

CHAPTER 3

A RELIABILIST ACCOUNT OF EVIDENCE

3.1 Introduction

Bayesian Epistemology sees rational belief change as spurred on by the receipt of evidence. This evidence is commonly taken to be a set of propositions of which the agent is certain. Beliefs are then updated by conditionalization on these evidence propositions. For some uses, we can simply stipulate what an agent's evidence is at various times. But for some purposes, we would like to know more about what it is for an agent to receive evidence. Any such account of an agent's evidence will be an addition to the standard Bayesian principles. But we would like an account of an agent's evidence to fit well with the existing structure of the approach. In the last chapter, I argued that it is consistent with Bayesian Epistemology to pursue an external account of what it is for an agent to have or receive some evidence. In this chapter, I will investigate a particular external account of evidence, and its advantages and disadvantages when paired with the rest of the Bayesian machinery.

The proposal that I will consider is a proposal that makes the having of evidence depend on the reliability of certain cognitive processes. Before beginning it is important to emphasize three things. First, though I will give and defend a particular account of having evidence, this is meant to be an *investigation*. Little has been said about what it is to have evidence.[1] The project I am engaged in can be seen as a detailed investigation of one way of sketching what it is to have evidence. Second, I will be investigating a

[1] Though this does not mean that *nothing* has been said. See, for instance: Field, ([1979]), Maher ([1996]), Williamson ([2000]), Silins ([2005]), and Neta ([2008]).

reliabilist account. Why reliabilism? I investigate this for two main reasons. First, it is one of several natural starting points: what makes some proposition evidence for you is that you are reliably related to its truth. Second, and more importantly, reliability is clearly an epistemically important property. When doing epistemology, we are often concerned with methods or processes of agents that maximize true beliefs and minimize false beliefs. Reliability is clearly relevant to this goal. Thus, if successful, a reliabilist account of evidence could also be *explanatory*: it could explain what is epistemically good about updating on one's evidence.

Finally, it is important to state that the account I give is an account meant for evaluators, not for agents. Now, each of us is both an epistemic agent and an epistemic evaluator, and theories or projects that help us to evaluate epistemic agents can also help us *be* better epistemic agents ourselves. But when this happens, the help is indirect. So, one should not look to the account that I will be giving for direct help as an epistemic agent. Instead, what we have is a tool for evaluation.

In this investigation, I will concern myself with conditionalization as an update rule. Conditionalization is usually understood as a sequential updating rule. Assume that time can be divided into discrete epistemic instants, moments at which epistemic changes happen. According to the traditional understanding of conditionalization, at each instant t, you update your credence function at $t - 1$ on the new evidence received at t. That is:

(sq-*COND*) $\mathrm{cr}_t(\bullet) = \mathrm{cr}_{t-1}(\bullet|E_t)$

where E_t is the new evidence received at t (or: between $t - 1$ and t). Since $cr_{t-1}(E_t|E_t)= 1$, it follows that $cr_t(E_t) = 1$.[2] Focusing attention on sq-*COND* is not insignificant here since it restricts us to evidence that is propositional and assigned credence 1. For those concerned with this restriction, I consider alternatives in Section 3.6.4.[3]

Given that we are concerned with sq-*COND*, an agent's evidence at a time will consist of a set of propositions. Call this the agent's *evidence set*. An account of an agent's evidence, then, will tell us how to pick the members of an agent's evidence set at a time. Of course, not just any set of propositions will do: the set of *all* propositions, for instance, is clearly inadequate. Not only is this intuitively absurd, it also renders an agent incoherent, since if all propositions are evidence then they all receive credence 1. Even restricting the evidence set to all *true* propositions is inadequate, since if all true propositions receive credence 1, we have rendered the interesting Bayesian machinery moot.

A different, and much more intuitively plausible account is:

> **Mental State Evidence (MSE)**: S's evidence at t is represented by the set of doxastic possibilities consistent with S being in the mental state, M, that S is in at t.

Let 'doxastic possibilities' mean the possibilities to which S does not assign 0 credence.[4] If we think of propositions as sets of possibilities, then according to MSE, S's evidence set contains at least the true proposition that S is in M. MSE gives one

[2] In the next chapter I will suggest an alternative to sq-*COND*, but for this chapter, I will adopt sq-*COND* as the correct way to understand conditionalization.
[3] Both Williamson [2000] and Neta [2008] gives arguments in favor of thinking that evidence is propositional. These arguments, however, are less than decisive. Nevertheless, thinking of evidence as propositional is certainly theoretically useful.
[4] Or possibilities to which S assigns a non-zero probability density.

kind of way of restricting the set of propositions to yield an agent's evidence set.[5] In contrast to this, I will consider something in the spirit of:

> **Reliabilist Evidence (RE)**: S's evidence at *t* is represented by the set of possibilities consistent with the propositions reliably indicated to S at *t*.

RE gives an alternative way of restricting the set of propositions to yield an agent's evidence set.

With this simple statement of RE we are in a position to see how it stands with respect to the issue of externalism. In his ([2005]) Nicholas Silins categorizes what internal and external accounts of evidence would look like in the following way:

> **Evidential Internalism (EI)**: Necessarily, if A and B are internal twins, then A and B have the same evidence.

The denial of this gives us:

> **Evidential Externalism (EE)**: Possibly, internal twins A and B have different evidence.[6]

What does it mean for A and B to be internal twins? Silins understands internal twins as two agents identical in all of their nonfactive mental states (e.g., having the same

[5] For the deployment of an account something like MSE, see Meacham ([2008]).

[6] How are EE and EI related to the various internalism theses considered in the last chapter? EE and EI do not themselves lay down epistemic requirements. Rather, they simply say what an agent's evidence *is*. However, if we pair EE or EI with sq-*COND*, then we do get an epistemic requirement. EE together with sq-*COND* will generate requirements that violate Access Internalism, Guidance Internalism, and Guidance Internalism*. Suppose two agents are internally identical and yet have different evidence. EE says this is possible and sq-*COND* says that these internally identical agents should update their belief states in different ways. If that's the case then the agents themselves will not have access to whether they updated correctly, since this access would require internal differences. Thus, Access Internalism is violated. Further, the agents themselves will nto be able to be guided by EE and sq-*COND* since it is internal features of agents that guide their epistemic behavior. Thus, Guidance Internalism is violated. Finally, since they are internally the same but have different requirements, it is possible that one but not the other is evaluated as doing something incorrect. Thus, Guidance Internalism* is violated.

beliefs, experiences, and apparent memories). Since Silins endorses EI, he appears committed to something in the spirit of MSE.[7]

Now, Silins makes clear that EI, as he means it, is *not* a thesis that requires agents to have internal *access* to their evidence. It is important to note that this is simply a consequence of his definition of 'internal twin'. The formulation of EE and EI allow different types of internalism/externalism, depending on how we understand 'internal twin'. We get different kinds of internalism about evidence, for instance, if we mean by 'internal twin' subjects who are identical with respect to: (i) what is inside the head, (ii) nonfactive mental states, (iii) doxastic states, (iv) what is cognitively accessible, (v) what is consciously accessible, etc. Some of these turn on access and some do not.

Given this, we can pose a question: does a reliabilist account of evidence (like RE) imply EE or EI? This depends on two factors. First, it depends on how we understand 'internal twin'; and second, it depends on whether such internal twins can differ in what is reliably indicated to them. It is likely that on most definitions of 'internal twin' RE will end up implying EE, and so be a form of evidential externalism.[8]

In Section 3.2 I will formulate a reliabilist account of evidence much more carefully. But even before giving the careful statement of the view, we have enough information to consider an objection to such an account. Considering this objection now will prove fruitful since it will help to clarify the kind of account of evidence being pursued.

---

[7] Which is not to say he endorses MSE. MSE tells you what evidence an agent has, EI tells you which kinds of agents have the same evidence, without necessarily telling you what that evidence is.

[8] It is perhaps worth noting, however, that RE *need* not do this if 'reliability' and 'internal twin' were understood in the appropriate way. So, though I have cleared space in Chapter 2 for external accounts of evidence, the reliabilist account I present here need not be developed in this way.

3.1.1 Neta's Counterexample: Clarifying the Project

In a recent paper, Ram Neta ([2008]) expresses pessimism over the ability to offer any kind of substantive account of what it is to have evidence. He considers a myriad of different proposals about what it is to have evidence, and offers counterexamples to all of them. In particular, Neta offers a counterexample to the view he calls "E=RB": *p* is a member of S's evidence set at *t* if and only if S's belief that *p* is reliably formed or sustained at *t*. (p. 103) This is a kind of reliabilist account of evidence. His criticism of E=RB appeals to a case where agent S is in neurological state N iff S believes that she is in N. Further, S is surprised to find she believes that she is in N. She just occasionally does believe it. S's belief that she is in N, then, is reliably formed. So, the proposition that she is in N is part of her evidence according to E=RB. But, says Neta, this is wrong. S is not required to distribute her confidence over hypotheses in proportion to the degree of support that those hypotheses receive from her belief that she is in N. That is, she is not required to conditionalize on the fact that she is in N. So, he concludes, E=RB fails.

Essentially Neta has constructed a Truetemp-style case (Lehrer [1990]), but for evidence. He then appeals to the intuition that such reliably caused beliefs are not the sorts of things we should use as evidence (or are required to use as evidence). The response to this case will clarify the kind of account evidence that something like RE is meant to be giving.

Notice first that there are two ways we can read the case. On the one hand, perhaps we are to imagine the case in such a way that S believes that that she is in N is *not* evidence. In that case, it is plausible that a reliabilist account of evidence has the

resources to handle the example. Indeed, in Chapter 5 I will argue that the reliabilist about evidence can offer an explanation for why, in general, agents shouldn't treat propositions as evidence if they strongly believe that such propositions are not evidence.

On the other hand, perhaps we are to imagine the case in such a way that S does not hold this sort of belief. Then, that she is in N *is* evidence according to a reliablist account, and it should be used as evidence. Is there anything wrong with this pronouncement? There is not so long as we distinguish what sense of 'should' is operative here. In his paper, Neta leverages two very different ideas about what evidence is. Each, I think corresponds to a different sense of 'should'. Once we are clear on this, we can see that the specific counterexample to E=RB loses its force.

Neta constructs counterexamples to many different accounts of evidence. The counterexamples can be divided into two specific kinds. The first kind of counterexample points to propositions that are labeled evidence by a certain account, but that are not connected with the truth in the correct way. For instance, Neta considers the proposal he calls "E=B": *p* is a member of S's evidence set at *t* if and only if *S* believes that *p* at *t*. This is objected to because one can form beliefs for no reason at all. According to E=B, the things so-believed are evidence. But, says Neta, it is clear that such propositions are not evidence.

The second kind of counterexample also criticizes proposed accounts for implying that an agent has too much evidence. But the nature of these counterexamples isn't that there is no connection with the truth, but rather that there is *too much* connection with the truth, which renders evidence unable to provide guidance to

agents.[9] For instance, Neta considers the proposal he calls "E=JTB": *p* is a member of S's evidence set at *t* if and only if S's belief that *p* is justified and true. This is objected to by imagining a situation in which the agent knows that 99 balls have been drawn from an urn, all of which have been black. Suppose it is true that the 100th ball to be drawn will be black. Before it is drawn, says Neta, the agent is justified in believing the 100th will be black. Since it is true, E=JTB says that the proposition that the 100th ball is black is part of the agent's evidence. If the 101st, 102nd, … draws are all black, then it appears that the agent has as evidence that all such balls are black. This, however, seems incorrect.[10] The complaint is that such propositions, though true, are not the sorts of things that should be *used* as evidence. But why not? I propose that the thinking behind this rests on the notion of guidance. For consider: even if the later draws *aren't* black, things will look exactly the same from the agent's perspective. But then any notion of evidence that says that such propositions *are* evidence will not allow evidence to guide the agent who is in such a situation.

These two desiderata, truth-connectivity and guidance, tend to pull in opposite directions. Neta's counterexamples nicely illustrate that these are two intuitive features that we'd like an account of evidence to have. It would surely be nice to have an account that completely captures *both* these features. But this seems unlikely: an account that focuses *exclusively* on evidence's connection with the world will say that an agent's evidence set contains all true propositions; an account that focuses *exclusively* on guidance will divorce an agent's evidence set from the world completely.

[9] It is worth noting that one proposal – the E=JB proposal, which identifies one's evidence with one's justified beliefs – receives an objection that doesn't seem to fit this two-category classification. Nevertheless, this does not undermine the general point made.

Both extremes are undesirable. Given that we are balancing two desiderata, it is certainly acceptable to focus towards one side or the other. Further, as noted in the last chapter, there is a kind of Bayesian model according to which we are not offering guidance to epistemic agents, nor are we attempting to construct epistemic robots. This approach naturally places less of an emphasis on guidance. The account of evidence that I will pursue follows this lead: it purchases some connection with the world at the cost of guidance. Reliability does just this, since it gets us a lot of connection with the world while still maintaining a connection to the agent's faculties and position in the world.[11] On this picture of things, it is a mistake to criticize an account of evidence for not always offering appropriate guidance. So, though it may be true that the agent in Neta's example should not use N as evidence if we are focused solely on guidance, it is certainly true that she should use N as evidence if we are focusing on the agent's connection to the world.

This distinction not only disarms Neta's counterexample to a reliabilist account of evidence, it also serves to clarify the project in showing what can reasonably be expected from an account of evidence. The term 'evidence', and more precisely, the phrase 'having evidence' can be understood in different ways, some of which are bound

---

[10] This proposed counterexample is very similar to an example that Williamson gives in his [2000]. I respond to this kind of objection in Chapter 4.

[11] There is an important point here. From Chapter 1, we are understanding Bayesian Epistemology as giving us evaluative, anti-procedural theories. This frees us from being completely constrained by the agent's computational abilities in our epistemic theorizing. In particular, it tells us that our theory need not detail a procedure the following of which allows an agent to conform with the theory. This does not mean, however, that we should reject any account that is sensitive to certain features of the agent in question. For example, we use credence functions to represent belief states because it seems as though, in certain respects, such credence functions are similar to the belief states that agents like us have. Here we have a constraint on our representation of belief states, based on the features of the agents we are trying to model. Thus, viewing BE as giving us evaluative, anti-procedural theories does not imply that there is something wrong with appealing to *some* features of the agents we are modelling. This is what a reliability-based account of evidence does. It takes account of certain features of agents (their reliable cognitive processes) to give an account of what their evidence is.

to be more valuable than others depending on the inquiry. The kind of account that I will offer is meant to give a precise statement of at least one salient sense of this phrase, and see how this account fits in with the machinery of Bayesian Epistemology. I do not, however, maintain that the account I will present will perfectly square with every idea of what it is to have evidence. Attempting to give an account that did that, I think, will ultimately prove unfruitful.

For example, some might want an account of having evidence with the following feature: in any situation, it is always obvious to the agent in question what evidence he has. This is not an unreasonable kind of thing to want from an account of having evidence. For instance, it would help to explain the plausibility of the claim that whatever else I might be ignorant of, I at least know what my evidence is. The account I will present, however, does not have this feature. But this alone does not show that the account is devoid of interest. I intend my account to appropriately say what evidence agents have in a variety of circumstances, including situations that arise in science and in the courtroom. However, the account is offered from an explicitly evaluative perspective. It will not, then, directly tell you what to do, or how to figure out "from the inside" what evidence you have. Instead, it offers an evaluative notion of having evidence, an account according to which evidence is a good thing to update on, and something we would ideally require of epistemic agents.[12]

With these preliminaries out of the way, we are ready to proceed to an investigation into a reliabilist account of evidence. A note about where we're headed will be helpful. I'll start off by stating a particular reliabilist account of evidence that I'll

call 'RAE'. This account will attempt to give necessary and sufficient conditions for an agent having evidence at a time. This chapter will be solely concerned with clarifying and explaining RAE and then responding to worries that arise after such clarifications have been made. In the next chapter (Chapter 4), I will consider more specific objections to RAE, as well as note scenarios in which RAE performs well.

3.2 Statement of RAE

In working our way to a statement of a reliablist account of evidence, we must consider a distinction. On the one hand, one might have an account of evidence that takes all evidence to be a subset of those propositions in which the agent has strong belief. Along these lines we might demand that no proposition is evidence for an agent at *t* unless the agent has a sufficiently high credence in that proposition (possibly = 1) at *t*. On the other hand, one might not require this: we might allow that a proposition that is not entertained or that receives low credence at *t* is nevertheless a member of the agent's evidence set.

This distinction maps roughly onto the distinction between doxastic and propositional justification. A proposition can be propositionally justified for S, even if S does not believe that proposition. Doxastic justification, however, is a kind of justification that only an actually held belief can enjoy.[13] Call accounts of evidence that

---

[12] A cautionary note: this does not mean that any purported counterexample can be dismissed as irrelevant. It is my contention that the account I sketch is still latching on to an interesting sense of what it is to have evidence.

[13] Peter Klein states the distinction succinctly:

> As the expression ''propositional justification'' implies, such justification is an epistemic property of propositions rather than a property of belief states. We can say that a proposition, h, is propositionally justified for S just in case there is an epistemically adequate basis for h that is available to S regardless of whether S believes that h, or whether S is aware that there is such a basis, or whether if S believes that h, then S believes h on that basis… A belief that h is

are similar to accounts of doxastic justification, *belief-dependent accounts of evidence*. Belief-dependent accounts maintain that one's evidence is a subset of what one believes, where 'belief' is taken to mean sufficiently high credence. Howson & Urbach hold that one's evidence consists of all and only the propositions to which one assigns full credence ([1993], p. 106). This is a belief-dependent account. Call accounts of evidence that do not require that evidence be believed, *belief-independent accounts of evidence*. I will begin by criticizing belief-dependent accounts of evidence. I will then propose a belief-independent account.

According to a belief-dependent account of evidence we could have two agents in identical situations, with their perceptual systems being stimulated in just the same way. Perhaps they are both looking at a red stoplight. One agent believes that there is a red stoplight in front of him, while the other does not. According to a belief-dependent account of evidence, this second agent has no evidence. This is implausible. There can be all sorts of things going on in front of the agent, his eyes can be wide open, he can be seeing lots of interesting things, he can be aware of all these things, and yet he doesn't change his doxastic state. Such an agent is, according to the Bayesian principles and a belief-dependent account of evidence acting in a perfectly rational way. But this doesn't *seem* rational. Belief-independent accounts have the edge here, as they need not say that such behavior is reasonable.

> doxastically justified for S when and only when S is acting in an epistemically responsible manner in believing that h. ([2007], 6)

For the same distinction see: Conee & Feldman ([1985]), who do not refer to the distinction by name; Kent Bach ([1985]), who refers to a person being justified versus a belief itself being justified; Alvin Goldman ([1979]), who refers to *ex post* and *ex ante* justification; Roderick Firth ([1978]), who refers to doxastic and propositional *warrant*; and Kvanvig & Menzel ([1990]).

Now, perhaps there is a kind of "thin" or deflationary notion of evidence according to which this is the right verdict. This verdict may even make sense if our aim is to provide guidance to an agent. Perhaps, as a rule of guidance, one shouldn't treat some proposition as evidence unless one strongly believes it. One might even argue for this via some sort of OIC principle. If 'ought' implies 'can', then 'not can' implies 'not ought'. Since it is not the case that one *can* treat a proposition as evidence if that proposition does not receive credence 1, it follows that it is not the case that one *ought* to treat such a proposition as evidence. But we have rejected OIC considerations, and if we are trying to say something substantive and evaluative about what it is to receive evidence, then this sort of account won't do. There is, it seems, something epistemically inappropriate about what the agent is doing in the case described above. Accordingly, it seems that a belief-dependent account of evidence is incorrect.[14]

It is worth pointing out here that a similar sort of example reveals a tension in Williamson's ([2000]) E = K thesis:

E=K: *P* is a member of S's evidence set at *t* iff S knows *P* at *t*.

Williamson takes knowledge as primitive and undefined. Nevertheless, in this context he tells us that knowledge entails belief. Given this, if *P* is part of S's evidence, then S believes *P*. But this means that if some agent does not have beliefs, then that agent has no evidence. So, imagine the following scenario. You draw a red ball out of an urn and show it to me clearly in bright light. I see the red ball. It seems to me that there is a red ball in front of me. Perhaps I even know that I am not in a skeptical scenario. Nevertheless, I do not believe that a red ball was drawn. Because of this, I do not know

[14] Neta [2008] reaches a similar verdict.

that a red ball was drawn, and so the proposition "a red ball was drawn" is not part of my evidence. But this seems wrong. My stubborn unwillingness to believe should not make this proposition fail to be part of my evidence set.

Williamson ([2000]) considers and responds to this sort of counterexample. He writes:

> According to E = K, my evidence includes *e* because I know that things are that way. But, a critic may suggest, that does not go back far enough; my evidence includes *e* because it is perceptually apparent to me that things are that way, whether or not I believe that they are that way […] The critic takes my evidence to be the evidence in my potential possession, not just the evidence in my actual possession. (pp. 202-3)

Williamson understands 'possession' here in terms of belief, and he argues for E=K based on the following example. Imagine that I am in the position to know all of the propositions $P_1, \ldots, P_n$, but that I actually only know P_k since I fail to believe the other propositions. Thus, I "actually possess" P_k but only "potentially possess" the other propositions. Now, consider a proposition Q that is very probable given all of $P_1, \ldots, P_n$, but improbable on any one. Writes Williamson: "According to the critic, Q is highly probable on my evidence. E = K gives the more plausible verdict, because the high probability of Q depends on an evidence set to which as a whole I have no access." ([2000], p. 203, notation slightly changed).

In response, one could claim that if I have no access to all the $P_1, \ldots, P_n$ as a whole, then it is misleading to say that I am in the position to know all of the propositions $P_1, \ldots, P_n$. Perhaps I am in the position to know *each* of the $P_1, \ldots, P_n$, but one need not conflate this with being in position to know *all* of $P_1, \ldots, P_n$. For it is very plausible to think that even though not *all* of the $P_1, \ldots, P_n$ are evidence in the case described, it is still true that *some* of them are evidence. And Williamson's account

can't explain this very basic fact. For, were I to believe *none* of them, then none of them would be part of my evidence. And this is clearly wrong. If we're aiming to give an evaluative account of what it is to have evidence, then believing *P* is not a necessary condition on *P* being evidence.[15] This reveals a tension in Williamson's E=K principle, since it is a belief-dependent account that is meant to be evaluative in nature. For such a project, belief-dependent accounts of evidence appear to be mistaken.

As mentioned above I am interested in investigating a reliabilist account of evidence. But traditional reliabilist accounts of justification are belief-dependent accounts. In the terminology introduced at the beginning of this section, they are accounts of doxastic justification, rather than propositional justification. Traditional reliabilism about justification, for example, says something like the following:

> **Reliabilism:** S's belief in *P* is justified iff it is caused by a reliable belief-forming process.

Despite this, there is a way of taking a doxastic theory of justification and turning it into a theory of propositional justification. For instance, Goldman ([1976]) offers the following:

> Person S is [propositionally] justified in believing *p* at *t* if and only if there is a reliable belief-forming operation available to S which is such that if S applied that operation to his total cognitive state at *t*, S would believe *p* at *t*-plus-delta and that belief would be [doxastically] justified.[16] (p. 351-2)

We can use something like this to give an account of evidence that is a reliabilist account, and yet is not a belief-dependent account. The core of the proposal that I will defend and investigate is thus:

---

[15] It's important to note that for all I've said, if something is evidence for you, it may be true that you *should* know it, or *should* believe it. In fact, this latter claim is a consequence of the view that I will present.

Reliabilist Account of Evidence (RAE): The set of propositions, $\mathbf{E_t}$ is S's evidence set at *t* iff there are reliable belief-forming processes available to S at *t* such that if S applied those operations S would believe all the members of $\mathbf{E_t}$ at *t* and those beliefs would be caused by the reliable belief-forming processes.[17]

Note that this escapes Williamson's counterexample. RAE does not have a consequence that all propositions $P_1, \ldots, P_n$ are evidence if the agent lacks access to all of them.The guiding idea behind RAE is that it is a high degree of reliability that makes a proposition evidence for an agent.

One plausible constraint on evidence is that the propositions in the evidence set be *true*. On reflection, it does sound rather odd to speak of *false* evidence. We routinely speak of *misleading* evidence, but this can and should be sharply distinguished from false evidence. For example, we readily assent that

G: there was a bloody glove at the crime scene

is misleading evidence, if we know that the glove was planted by corrupt police officers. This is because, given usual sorts of background beliefs, *G* will lead one to conclusions that are incorrect. Nevertheless, *G* itself is still true: there *was* a bloody glove at the crime scene. Once we distinguish misleading evidence from false evidence, intuitively plausible cases of false evidence are hard to find. Thus, our account should deliver the result that, at least most of the time, the propositions in an agent's evidence set are true.

---

[16] Goldman actually uses the terms 'ex post' and 'ex ante' justification, but the intention is the same.
[17] This account bears some similarity to a proposal made in an unpublished (but recently posted online) paper by Brian Weatherson ([*ms*]). In the paper, Weatherson defends an account of evidence according to which one's evidence set includes the propositions that are the outputs of reliable input modules (in the sense of Fodor ([1983]). Weatherson's description of his idea is short, but the account clearly has similarities to RAE. There are, however, important differences. In particular, RAE and Weatherson's

RAE has this result. Since the propositions in an agent's evidence set are reliably indicated, most of them are true. However, some might find it plausible that evidence *must* be true. If so, one could defend a variant of RAE that adds the constraint that the members of $\mathbf{E}_t$ must be true. Call such an account 'RAE-t'.

It is an interesting question whether or not RAE or RAE-t do a better job at giving us an account of evidence. Note that if we require *perfect* reliability, then the truth condition in RAE-t is superfluous. So, assuming a high level of reliability, RAE and RAE-t overlap in a majority of situations. But there are cases where they come apart. The issues that will concern us in this chapter and in the next will not do much to distinguish RAE from RAE-t. However, in later chapters the truth condition will make something of a difference. Accordingly, I will note when situations arise that would distinguish RAE from RAE-t. However, when it won't cause confusion, I'll simply use the term 'RAE', letting this be shorthand for both RAE and RAE-t.

Above I noted that an account of evidence will give some way of selecting a privileged subset of the propositions as the agent's evidence. RAE does this by restricting an agent's evidence to those propositions to which the agent has a reliable route. RAE-t does this by restricting an agent's evidence to those *true* propositions to which the agent has a reliable route. Note, too, that both RAE and RAE-t are forms of Evidential Externalism if we grant that internal twins can differ in the reliability of the belief-forming processes that they employ.

With the view stated, I will now respond to several natural questions that arise. In responding to these complications, RAE will be further clarified.

---

account disagree about the status of *inferential processes* of belief formation, as well as the role that background beliefs can play in evidence acquisition. Chapter 4 takes up these issues in more detail.

3.3 Belief Versus Degree of Belief

The guiding idea behind RAE is that what it is to have some proposition as evidence is to have a reliable route between you and the truth of that proposition. If we formulate the idea in that way, however, 'you' is ambiguous. RAE is an attempt to make this guiding idea more precise. The reliable route must be between S's *doxastic state* and the truth of the proposition. One might legitimately worry here, however. For we are conducting this investigation within the confines of Bayesian Epistemology. Bayesian Epistemology countenances doxastic states that consist of *degrees* of belief, not all-or-nothing belief. However, RAE seems to be formulated purely in terms of all-or-nothing belief. How, then, can RAE give us an account of evidence for Bayesian Epistemology?

One way of going is to change RAE so that the core clause instead reads:

> **RAE-full:** The set of propositions, $\mathbf{E_t}$ is S's evidence set at *t* iff there are reliable processes available to S at *t* such that if S applied those operations S would have **full credence** in all the members of $\mathbf{E_t}$ at *t* and **full credence** in those propositions would be caused by the reliable belief-forming processes

In some ways this alteration may be acceptable. For if RAE is accepted as an account of evidence, then it will follow from *COND* that all those reliably caused beliefs *should* receive full credence. And, in some cases, we simply stipulate that agents have full credence in some propositions which are their evidence. In these cases, RAE-full will work fine. But we also want an account of evidence to offer us (as evaluators) guidance about how to describe various cases where it *isn't* stipulated what the agent's evidence is or that the agent gives such evidence full credence. I worry that in such situations

RAE-full is actually less clear than RAE. This is simply because while we have a good sense of what reliable belief-forming processes are, we might not have a very good sense of which processes are reliable full-credence producing processes, since 'full-credence' is a philosophical term of art.

A better way of going, I think, is to stick with RAE but think more carefully about what we're doing with Bayesian Epistemology. What we want from an evaluative theory is the ability to evaluate the epistemic performance of agents like us. It is undeniably correct to say that agents like us believe certain things in an all-or-nothing sense. Perhaps this can be translated into degree-of-belief talk in terms of some threshold. For instance, perhaps all degrees of belief greater than 0.9 count as all-or-nothing beliefs. But it is likely that there is no generally applicable, simple translation from degree-of-belief talk to all-or-nothing belief in this way.[18] Despite this, we *do* understand all-or-nothing belief attributions; we understand what it means to say that my belief that *P* is caused by a reliable belief-forming process, and we understand what it means to say that I have a reliable belief-forming process that would have resulted in the belief that *P*. We can understand this even if we cannot perfectly translate this all-or-nothing belief talk into degree-of-belief talk. The proposal, then, is to consider S's reliably formed beliefs and those not-believed propositions to which S has a reliable route. These propositions form S's evidence set. Accordingly, in the evaluative Bayesian representation of this agent, such propositions constitute the agent's evidence

[18] The Lottery Paradox (Kyburg [1961]) and the Preface Paradox (Makinson [1965]) are particularly clear examples of why this is difficult. For further discussion of this issue see Christensen ([2004]), Folely ([2009]), and Frankish ([2009]).

and thus *should* receive credence 1.[19] In saying this, I do not mean to denigrate the interesting subject of the relation between degrees of belief and all-or-nothing belief. Rather, I simply point out that we can understand what it means to have a process that produces a belief even if we do not have a perfectly worked out theory about the relationship between degrees of belief and all-or-nothing belief.

3.4 Availability

According to RAE my evidence depends on the reliable belief-forming processes that I have available to me. So, just what belief-forming processes *are* available to me? For instance, consider a series of cases:

> *Case 1*
>
> I am sitting in my office wondering what the temperature is outside. My properly working eyes are directed at the accurate thermometer that is outside my window. In this case it seems as though I have available to me a reliable belief-forming process that yields the belief that it is 54°F outside.

> *Case 2*
>
> I am sitting in my office wondering what the temperature is outside. My properly working eyes are directed slightly away from the accurate thermometer that is outside my window. By shifting my gaze slightly, I could bring the thermometer's reading into my visual field. In this case, too, it seems as though I

[19] If there is some accepted translation between all-or-nothing belief and degrees of belief, then I would endorse the translated version of RAE over RAE. In particular it is important to note that I do not intend RAE to commit one to a distinct kind of doxastic state (all-or-nothing) to stand in contrast to degrees of belief.

have a reliable belief-forming process available to me that yields the belief that it is 54°F outside.

Case 3

I am sitting in my office wondering what the temperature is outside. There is no thermometer outside my window. But if I go down the hall to my colleagues office, I can quickly find out what the temperature is by looking at his thermometer. In this case, is it true that I have a reliable belief-forming process available to me that yields the belief that it is 54°F outside? It is less clear.

We could easily multiply cases like this so that there are different situations where a certain reliable process is available to the agent in differing degrees. What this shows, I think, is an important kind of context-sensitivity in our notion of *having* evidence. Depending on what we want to model, and the context of the inquiry, we can allow the notion of availability to vary as needed.

For instance, consider a juror who has been nodding off in a courtroom. If we restrict the processes available to the juror to those processes the juror actually used, very few true propositions about the case are available to the juror. But it is not at all odd to say that the juror's evidence at least includes propositions about what was relayed during the trial. This is especially true if we are offering an evaluative epistemic theory. Given the situation, this evidence was available to the juror, despite the fact that the juror did not actually make use of that evidence. We could plausibly say that the juror should believe that evidence, and adjust his other beliefs accordingly. In this case,

it seems that which processes are *available*, is determined by things that jurors *should* do.

Now, one could respond to such a case by claiming that the juror only had the evidence that was available to him, given the processes he actually engaged in. On this way of understanding 'available' the juror's evidence is minimal. One could then point out that the juror made a kind of error in not being in a state in which he could gather more evidence. This is a different way of modeling the case. It seems to me that both senses of 'available' give acceptable, useful, and understandable accounts of the evidence that the juror had, and should have responded to. It thus seems appropriate to allow the notion of availability to vary as needed.

There will, however, be more or less standard contexts. For instance, it is plausible that one standard context is where we fix the term 'available' by fixing the orientation of the agent's perceptual devices. This will give us the result that only in Case 1 does the agent's evidence include the proposition about the temperature. Similarly, we might even fix the term 'available' by saying that the available processes were only those that the agent actually used. In this case, RAE can mimic belief-dependent accounts of evidence.

RAE, then, delivers different verdicts about cases depending on how availability is understood. If there is *one* way to understand that term that is appropriate to epistemic evaluation, then RAE should adopt this. But it doesn't seem as if this is the situation. The variability in RAE seems to be a virtue of the account.

3.5 The Generality Problem

Here's a way of figuring out what an agent's evidence is, according to RAE: First, we list all the propositions (or, if we are using RAE-t, we list all the *true* propositions). Then, we ask which set of propositions is such that the agent has a reliable way to believe each of those propositions. To answer this question, we need to know something about the *ways* that an agent can believe various propositions. This brings up the much-discussed generality problem for reliabilism.[20] RAE says that the evidence an agent has depends on the reliable processes of belief-formation the agent has available. But what sorts of processes do we appeal to here? For consider the following scenario. I am looking at the green blackboard in my office in normal conditions. Consider the proposition:

GB: The blackboard is green.

If I believe *GB*, is *GB* formed by a reliable process? If I do not believe *GB*, is *GB* such that there is a reliable process that would result in my believing *GB*? These questions can seem to have no answers, for there are various ways of specifying the ways that I could come to believe *GB*. Consider one extreme: my way of believing *GB* (the belief-forming process) is just the way I believed at this specific time, in this specific situation. Thus, that way (that process) will be reliable just in case the blackboard is green, and unreliable if not. Consider the other extreme: my way of believing *GB* is specified completely generally, say, as believing a proposition. Thus, that way (that process) will almost surely be unreliable. Consider a middle position: my way of believing *GB* (the belief-forming process) is specified as believing propositions on the basis of human

[20] Goldman ([1976]) makes reference to this problem, but it is Feldman & Conee ([1985]) that really made the problem stick.

visual perception in good environmental conditions. This, presumably, will give a different reliability score than either of the extremes.

The generality problem is the problem of specifying the preferred way of describing the process that led (or could lead) to my belief that *GB*. Since different descriptions lead to different reliability scores, it is important to answer this question. RAE, of course, is compatible with many different ways of specifying *the* process of belief-formation. If the processes are specified in different ways, RAE will give different results. Accordingly, RAE can leave this unspecified and allow the way of specifying the processes to be determined as needed. In many cases it is clear what the relevant process is; in some cases it is not. It may be foolish to search for one criterion that can answer the generality problem once and for all. We may not have *one* generality problem but instead many, local generality problems, each requiring their own solutions.

Nevertheless, if the defender of RAE can say *nothing* about the generality problem, it threatens to render RAE uninteresting, as it cannot be applied to any cases. Thus, I will here argue that the generality problem is not a hopeless one, relying on various solutions have been offered. Several of these solutions seem promising, and as far as I can see, the problem raises no special issues for RAE. Further, even if no solution so far given is adequate, I am convinced that the generality problem is a problem for everyone, and so it is not incumbent on a reliabilist—specifically a reliabilist about evidence—to decisively answer the worry.

One line of response to the generality problem is given by Mark Heller ([1995]). He argues for a contextualist solution. In different contexts, different processes are

salient. If adopted here, this would have the result that what is in your evidence set is context-sensitive. Though not my preferred solution, this is not obviously absurd. I argued above that our thoughts about evidence are context-sensitive in a different way. Adopting Heller's proposal, would simply introduce a new dimension of contextual sensitivity.

A different and promising line of response is given by James Beebe ([2004]). Beebe argues that the reliabilist can whittle down the eligible belief-forming processes in a principled way. He does this in two steps. We can avoid the hyper-generality extreme (the process by which I formed the belief was simply the process of forming a belief) by appealing to what he calls the tri-level condition. This tells us that T is a relevant process type only if all members of T

(i) solve the same type of information-processing problem,

(ii) use the same procedure or algorithm to do this, and

(iii) share the same cognitive architecture. (Beebe [2004], p 180)

To see how this might work, consider the specific process (T1) of forming a belief about what is in front of one via one's visual system in good light, and the more general process (T2) of forming a belief. Consider a token process that is a member of both T1 and T2. The tri-level condition tells us that the relevant process relative to which the token is to be assessed for reliability is *not* T2. For there are members of T2 that do not meet (i) and (ii). (It is an open question whether or not T1 meets the tri-level condition.) This shows how we can whittle down the class of competing process types. The basic idea is that we can appeal to cognitive science and psychology to help us locate the kinds of natural processes of which certain belief-forming tokens are instances.

The second step is to avoid the hyper-specificity in specifying process types. Beebe does this by appealing to Wesley Salmon's statistical relevance condition. The basic idea is that the type of process that we appeal to must be the broadest objectively homogenous subclass of the type given by the tri-level condition that subsumes the token process in question. Put simply, this requires that the token is placed within the most general type such that there are no statistically significant subclasses into which that type could be subdivided (Beebe [2004], pp. 187-90).

For example, imagine that on Monday and on Tuesday I looked on weather.com to see if it would rain. On each day, as a result of that day's token process, I formed the belief that it would rain. I might try to classify these tokens as separate process types: the process that leads me to believe that it will rain as a result of looking at weather.com on Mondays, and the process that led me to believe it will rain as a result of looking at weather.com on Tuesdays. Let 'W' refer to the general process type of forming a weather belief based on the weather.com forecast, and let 'WM' and 'WT' refer to the more specific processe types relativized to days of the week. Let *p* refer to some token belief-forming process. It is plausible that:

$$P(p \text{ produces true belief} \mid p \in WM) = P(p \text{ produces true belief} \mid p \in WT) = P(p \text{ produces true belief} \mid p \in W).$$

This shows that the processes WM and WT are not the broadest objectively homogenous processes that t_M or t_T belong to. Accordingly, those more specific processes are not the relevant types for assessing reliability in this instance.

One might wonder what happens when these are not equal. For instance, consider a situation where the subclass in question is "W on *that particular* Monday", and on that particular Monday it did rain. Then it might seem that

P(*p* produces true belief | $p \in$ W on *that particular* Monday) = 1

whereas

P(*p* produces true belief | $p \in$ W) < 1

Thus, it would seem that the broadest objectively homogenous process that t_M belongs to is the maximally specific process "W on *that particular* Monday", and this is not what we want.

We can, however, sidestep this problem. For we must distinguish what is meant by the process "W on *that particular* Mon". If the only member of "W on *that particular* Mon" is t_M then indeed the relevant probability is 1. But it is a plausible constraint on the relevant process type that no relevant type consists of only one token. If that's the case, then "W on *that particular* Mon" must refer to a type of process with many distinct tokens, occurring at different worlds. If this is the case and weather.com works in anything like the way we think it does, then:

P(*p* produces true belief | $p \in$ W on *that particular* Monday) =

P(*p* produces true belief | $p \in$ W)

Now, it might be that these probabilities are different. It might be that something about weather.com made it a more reliable source on that *particular* Monday. (Perhaps an all-star weather forecaster was working that day, but then quit before giving any further forecasts.) If so, then the reliabilist can happily grant that this *is* a relevant process of

belief-formation, and that the day of the week can make a difference to the epistemic status of a belief so-produced.[21]

Given what has been said, there are still ways of trivializing the way in which processes are typed. Consider, for instance, the process: "Believing weather.com's prediction of rain when it will rain." This is not a trivial type of process because there is more than one token in the class, and yet it seems to be a homogenous subclass of the process: "Believing weather.com's prediction of rain." So, Beebe's solution still seems to allow this trivializing way of typing the processes of belief formation. What can be said in response?

To some extent, what can be said depends on what is wanted as a response to the generality problem. If what we want is simply some way of picking out the relevant type that is intuitively correct and does not do so in a case-by-case way, then there is a simple response. The response is to say that we cannot appeal to the truth of the belief in typing the belief-forming process.[22] By so-appealing we build reliability information into the process too directly, and so get unintuitive ways of typing processes. Adding this stipulation will get around this problem and others like it. Further, one can give this response before being presented with the case at hand, so it is not simply a case-by-case fix.

However, some might want more from a solution to the generality problem. Some might be unhappy with such stipulation, and instead want an explanation for such a constraint. Here's what one might say to motivate the constraint proposed. When we

[21] Of course, one might not like this particular feature of reliabilism, but it is important to note that *this* feature has nothing to do with the generality problem. Rather, it is a feature of externalism, in general.
[22] A more general constraint might say that we can't appeal to the probability of the truth of the belief in typing the belief-forming process.

appeal to the truth of the belief in typing the process, we appeal to something that is no part of what the agent is sensitive to. In the case considered, the agent is sensitive to different features of her proximal environment, including the weather.com report, but not to whether or not it actually *will* rain or not. Given this, it is inappropriate to type belief forming processes in this way.[23] This is a way of motivating the constraint given one paragraph above.

Now, one thing to note about this motivation is that it would seem to rule out certain forms of externalism. For instance, consider a form of externalism that tries to distinguish between visual beliefs formed by regular perception under good lighting conditions, and visual beliefs formed by regular perception under misleading lighting conditions. Since the agent is not sensitive to the lighting conditions, this proposal will say that we cannot distinguish these two processes of belief formation. If we go for this motivation, we will have a more internalist externalism. We don't *entirely* lose the externalism, of course. Depending on whether poor lighting is exceptional or the norm, the reliability of regular visual perception may change. Further, even with this admission to internalism, any view like the one that I am sketching will be externalist in the sense that the agent certainly doesn't have access to whether or not a given proposition is reliably indicated or not.

Nevertheless, it is worth noting that offering the kind of motivation I just did for typing belief-forming processes will move things towards internalism. Some externalists may balk at this admission. Given my project, however, this is not a worry.

[23] One might even use the tri-level condition to help here. That is, one might say that any two processes that are tri-level identical must be part of the same subclass. This may or may not help depending on how we understand 'information processing problem'.

The point of the project I am engaged in is not to defend an externalist account of having evidence. Instead, the project is to give and investigate a *reliabilist* account of having evidence. I defend externalism because reliability-based accounts end up having externalist features. This brings up an important methodological point. Reliability considerations are capable of doing real epistemological work since reliability is clearly epistemically important. Internal/external considerations, on the other hand, do not have such important epistemic implications. The internality/externality of a given feature is not directly epistemologically relevant. If, driven by reliability considerations, that account ends up looking somewhat internalist, then this is no problem. If we are pushed to internalism by reliabilism, then this seems to be a good reason to be an internalist.

In summary, then, adopting something like Beebe's proposal seems to offer a promising resolution to the generality problem, and one that the proponent of RAE can adopt. We use the tri-level condition and the statistical relevance condition to narrow down the class of relevant process types. Then, we maintain that no process type can include only *one* token, and that no process can be typed in a way that appeals directly to the truth of the belief in question or the probability that the belief in question is true. Note that even if this proposal does not leave us with *one* distinct process type, it will still reduce the candidates significantly. Given this, one might pursue a Heller-Beebe hybrid response in which we first whittle down the acceptable processes and then go contextualist for any remaining processes.

3.6 How Reliable?

A third kind of question about RAE asks *how* reliable a process must be before it is said to be reliable enough to yield evidence. To answer this, we first must say something

about how to think about reliability. First we take the belief-forming process and use this to come up with a set of belief-forming episodes that are instances of that process. This set could consist of only actual instances of the process, or could consist of actual and non-actual instances of that process. After this, we must come up with the level of reliability necessary for a process to count as giving an agent evidence. Let the reliability of the belief-forming process be represented by *r* where *r* = the number of episodes in the set where the belief is true, divided by the total number of episodes in the set.[24]

Now that we have defined *r*, we must say what level *r* must meet for a process to be reliable enough to yield evidence. One option is to say that for a process to give an agent evidence *r* must = 1. This, however, seems overly restrictive, for it is doubtful that any belief-forming process is perfectly reliable without some severe gerrymandering of the sets used to assess reliability.[25] Thus, we should say that a process is reliable enough to yield evidence when $r \geq t$, for some threshold *t*.

So, just what is this threshold? Again, it seems right to allow variability in what counts as evidence, and so grant that there is not one fixed value of *r* that is required. Consider: we want to know what a juror's evidence is in a certain case. It seems correct to say that the juror's evidence includes propositions describing the crime scene that were relayed in court by the investigating police officer. But, of course, if we start wondering whether or not the officer was lying, we won't want to say that propositions describing the crime scene are evidence for the juror. Instead, the fact that the officer

[24] If there are infinitely many members of the set, then this procedure will not work. When working with infinite probability spaces we will have to introduce some reasonable measure over the space. How precisely to do this is a difficult question.
[25] However, see Lewis's "Elusive Knowledge" in his ([1999]) for thoughts on this.

reported those facts are evidence. This can be true *even if* the juror himself doesn't have any beliefs about the police officer's veracity. Further, if we then start wondering whether or not the juror is in the Matrix or not, then we won't want to say that the officer's reports are part of the juror's evidence. Instead, the fact that it seemed as though the officer reported those facts are the evidence. What we have here, it seems, is a shift of context. By allowing a variable threshold we can model these different situations.

Bringing out this feature of RAE allows us to see what kind of constraint on evidence RAE is. In essence, what RAE is doing is giving us certain classes of permissible evidence sets. That is, RAE tells us that if P is evidence (and P is reliably indicated at level r), then every other proposition indicated at level $\geq r$ is evidence and there are no evidence propositions indicated at level $< r$. In this way, RAE helps spell out various commitments we have about evidence sets, given that we want to include certain propositions in the evidence set. Thus, one should not see RAE as telling you what an agent's evidence is at a time, full stop. Instead, it provides something like a consistency criterion for what an agent's evidence is at a time. The proclamations that the account will give take the form: *given that P is evidence at t, so are* $\{E_i\}$.

One might worry that this renders RAE unfalsifiable. Any counterexample is given the response: make a new model! No counterintuitive situations seem to reflect badly on RAE. It is important to point out that this is not the case. There are situations where RAE would clearly be falsified. These are cases where, in one context, some proposition indicated at reliability level r is clearly evidence, and yet some proposition indicated at level $n > r$ is clearly not evidence. Such a situation is one that would tell

against RAE. In fact, in Chapter 4 we will encounter lottery scenarios that have just this structure to them, and which motivate modifications. So, this way of understanding of RAE does not render it vacuous.

Let me summarize how RAE has been clarified so far. First, RAE appeals to belief-forming processes, but this does not indicate a commitment to two distinct kinds of doxastic state: all-or-nothing belief and degrees of belief. Second, RAE has some variability in what processes of belief formation are taken to be *available*. By changing what is meant by this, RAE will give different verdicts. Third, RAE is not meant to have variability in *how* we type the processes of belief-formation. I have given a response to the generality problem that is supposed to remove this variability. Fourth, RAE allows less-than-perfectly reliable belief-forming processes to yield evidence. Finally, according to RAE, the level of the reliability threshold for a process to yield evidence is variable.

I anticipate that the last of these two features of RAE may lead to concern. In what follows I will formulate and respond to the most serious of these worries.

3.6.1 Variable Threshold Worries

3.6.1.1 The "Real" Evidence

Above I gave a courtroom example where a context-shift alters what propositions are naturally described as being part of the agent's evidence. As we evaluators entertained more and more skeptical scenarios, the reliability threshold increased until only facts about the juror's seemings ended up as evidence. One might object to this variable threshold approach, however, and instead simply say that the juror's evidence all along

consisted of various facts about how things seemed to her. This is her *real* evidence, the thought goes. According to this view, there is no need for a variable threshold.

I don't think that this is the correct response to such scenarios for two reasons. First, these sort of shifts to the agent's "real" evidence, seem to have no stopping point. For just as it can seem inappropriate to treat the officer's reports as evidence, so too can it seem inappropriate to treat the way things seemed as evidence.[26] Second, insisting that the juror's evidence all along consisted of various facts about how things seemed to her seems severely unmotivated. For what is so epistemically special about how things *seem* to an agent? One promising answer is that agent's typically have higher-than-usual reliable routes to how things seem to them. If that's right, then by setting the degree of reliability very high, we would have a principled reason for having an agent's evidence consist of how things seem to that agent. But note that if we do this, then there is no principled reason why there couldn't be evidence that doesn't go via *seemings* in this way. Reliability is what is driving the account of evidence here. And once we see that, we might wonder why we couldn't have non-seemings be evidence, or why (in certain situations) a lower reliability threshold isn't appropriate. Allowing the level of reliability to vary allows us the ability to countenance different conceptions of evidence by allowing context to shift the level of reliability that must obtain for some proposition to be evidence.[27]

---

[26] For more on this see the next section and Section 4.3.3 (in Chapter 4).

[27] Note that when we increase the degree of reliability that we require for something to be evidence, this has the effect of giving us a picture of evidence that looks very much like an EI account. This is because of the contingent fact that often we are more reliable with respect to propositions about what mental state we are in rather than propositions about the external world. It is very important to note, however, that such an account is not an internalist account. Various propositions about the agent's mental state are evidence because of facts about reliability, not because of their internality. Thus, even if there are situations where RAE and an internalist account of evidence give the same verdict, they do it for different reasons. These differences will come out in counterfactual scenarios.

3.6.1.2 Dependence on Evaluator

One might have a different kind of worry with the variable threshold approach: what is worrying is not that the threshold is variable, but rather than its variability depends on the evaluators, and not on the agents themselves. One might legitimately wonder why varying the evaluator's interests would change what is evidence for the agent, and thus to what degree it is reasonable for the agent to believe various propositions.

Notice that behind this kind of objection to RAE seems to be a kind of supervenience claim: If two agents are in identical situations, then what it is reasonable for them to believe is the same. RAE seems to violate this. For we can have two agents in identical situations (both internally and externally), but where one agent is evaluated according to one threshold, and the other agent is evaluated according to a different threshold. Different credences would be reasonable for each agent. Now, this way of putting things is somewhat contentious. For the defender of RAE *does* agree that the same credences are reasonable for two agents in identical evidential situations (with the same prior credence functions). So, the defender of RAE can say that in the objectionable situation described above, the two agents' situations are *not* identical: since the reliability threshold is different for each of them, their evidence is different. But perhaps the objector could modify the intuitive supervenience claim that underwrites the objection. Perhaps the driving intuition is that if two agents are in

---

For example, imagine a creature that has very reliable access to what kind of object is in front of it, but unreliable access to his own sensations. Imagine that this creature always assumes that if it believes that there is a tree in front of it, then it must visually appear as if there is a tree, but that this is incorrect. It is normally some other sense modality that indicates the presence of the tree. RAE will not say that propositions about those appearances are evidence, although the proposition about the presence of the tree could be evidence.

physically identical situations, then what is reasonable for them to believe must be the same. RAE does not seem to respect this claim.

What should the defender of RAE say? I think there are two responses. First, it is not unusual to evaluate two identical situations differently, *given* two different standards of evaluation. For instance, imagine that a paper I receive from an Introduction to Philosophy class is identical to a paper that I receive to referee on behalf of *Journal of Philosophy*. There is nothing odd with me evaluating these identical papers differently, given that there are different standards of evaluation. The defender of RAE can claim that he is doing something similar. The defender of RAE *never* has to say that he evaluates two physically identical agents differently *according to the same standard of evaluation*.

The second response is to acknowledge that, to some extent, we do have the intuition that physically identical situations should yield identical epistemic evaluation, but note that there might not be an account that could deliver on such an intuition. For think about a certain situation, described in as much physical detail as possible. What *are* the reasonable degrees of belief for the agent to have in that situation? To answer this, if we're Bayesians, we will have to specify the agent's evidence in that situation. In that situation there either will or will not be some propositions that the agent can infallibly detect the truth of. If we take seriously the idea that no propositions are like this (that is, if we take fallibilism seriously), then for any evidence set we consider we can legitimately question if that set *is* the agent's evidence set. So, we'll either get pronouncements about the agent's evidence set that can legitimately be questioned, or we'll get the extreme pronouncement that the agent has *no* evidence. If we rule out this

extreme pronouncement as unacceptable, then any evidence set we specify will be extremely unstable from the perspective of the evaluators, in that it is always possible for us to legitimately question some proposition's membership in the evidence set. But then it looks like we're never going to get just *one* answer to what the reasonable degrees of belief are in a situation. The best we can do is to give several answers, in some sort of principled way. This is what RAE attempts to do. It helps evaluators tell coherent stories about what an agent's evidence is. It doesn't say that these are the *only* stories that can be told. But this is no slight against RAE since it's not obvious that there *is* just one story to be told. It is no slight even if it is obvious that there is one story to be told, but not obvious what that story is. In such a situation, RAE can serve as a useful tool.

Now, this whole line of argument might suggest a different kind of objection to RAE. One might think that this presents us with a good reason to think that all that matters with respect to a proposition getting in an agent's evidence set is the agent's degree of certainty in a proposition, not actual data about the agent's fallibility with respect to believing that proposition. This view resolves the issue nicely in that an agent's evidence set contains all and only the propositions of which the agent is fully certain. Since which propositions the agent is fully certain of is part of the physical description of the situation, the degrees of belief that are reasonable for agents in physically identical situations are the same. However, this kind of view comes at a cost, since it is a deflationary view of evidence. According to such an account, full confidence in a proposition renders it evidence, no matter what the proposition, or the agent's relationship to the truth or falsity of that proposition. If we want to give a more

substantive account of evidence (and I've already argued that we do), we can't adopt this kind of view.

There is, however, a still different kind of objection to RAE. Above I argued that if we are epistemological fallibilists, then we're unlikely to get one definite answer from a physical specification of the situation, about the reasonable degrees of belief. But this argument was based on the idea that we are working with a model where evidence gets credence 1. One might think that it is plausible that there *will* be one story about what the reasonable credences are for an agent when we allow evidence that comes in degrees. In Section 3.6.4 I will address this issue head-on. But for the moment let us assume that there is some account, according to which evidence comes in degrees, and according to which we can specify what the reasonable degree of belief is for the agent, given only a physical specification of the situation. Even if we had all this, RAE could still be of use as a tool for evaluators. Degreed evidence brings with it quite a bit more complexity. RAE would then be useful as a more fecund way to deal with epistemic situations in a somewhat idealized way.[28]

3.6.2 Less Than Maximal Reliability Worries

In the previous section I considered worries with RAE based on the fact that RAE employs a variable reliability threshold for evidence. In this section, I will consider worries with RAE based on the idea that, according to RAE, less-than-maximally reliable processes can yield evidence that receives credence 1.

[28] Perhaps it is worth noting that the core of RAE is intact even if it is not the evaluator who sets the reliability threshold. It could be some feature of the agent's situation that sets this threshold. Now, I do not know what this feature would be, so I have not advocated such a view, but the general idea behind RAE is not inconsistent with such a modification.

3.6.2.1 Unrevisable Credences

One worry in this vein is that a less-than-maximally reliable process shouldn't be able to give an agent full credence in a proposition, since full credences are unrevisable. If sq-*COND* is our update rule, then once a proposition is assigned credence 1, it stays there forever since if cr(E) = 1, cr(E|•) = 1 (when defined). This is a legitimate worry, but in the next chapter (Chapter 4) I'll argue that this is not a consequence of RAE if we understand Conditionalization in a different way, which allows agents to lose evidence.

3.6.2.2 Betting Considerations

For largely historical reasons, there has been thought to be a tight connection between degrees of belief and betting odds that the agent thinks to be reasonable. Intuitively, if cr(Heads) = 0.5, then one takes it to be reasonable to bet on heads at 1:1 odds. Some, in fact, understand a degree of belief in a proposition as simply a way of encoding the odds at which an agent would be reasonable to bet on that proposition's truth. If there is this tight connection between betting and degrees of belief, then we can pose problems for an account like RAE. First, according to RAE, a less-than-fully reliable process can give an agent some proposition as evidence, mandating credence 1 in that proposition. But this means the agent would be reasonable to bet at *any* odds on the truth of that proposition—something that seems suspect if the process is not fully reliable. Further, according to RAE, what evidence an agent has can change depending on the level of the reliability threshold, which is determined by the evaluators. Thus, the evaluators appear able to change what odds it would be reasonable for an agent to bet at. This, too, seems suspect.

The correct response to this challenge, I believe, is to downplay the importance of betting considerations. Although historically degrees of belief have been associated with reasonable betting odds, this association deserves to be questioned here. In particular, a negative answer to the question, 'Would you bet on that proposition at those odds?' is not sufficient to show that such a degree of belief is unreasonable, at least in the context of our inquiry. The reason for this is that considerations of whether or not a bet is considered fair concern issues very much related to guidance and things internally accessible. After all, I might not bet on a proposition at certain odds, even if my degree of belief in that proposition corresponds to such a bet, if I don't *know* what my degree of belief is, or if I don't have access to this fact about myself. Similarly, it might not be reasonable for me to bet at *any* odds on certain logical truths, even though the probability calculus demands that I give such propositions credence 1. In cases like this we see considerations of reasonable betting odds pull away from considerations about reasonable credences. In short, reasonable betting odds seem to be highly influenced by what the agent has available to her, and how things seem from her perspective. If we are giving a third-personal, evaluative account of having evidence, however, then considerations about accessibility, guidance, and first-person perspective are not at the forefront. Thus, we should not expect reasonable betting behavior to track reasonable credences in this context.

So, the defender of RAE can legitimately put aside betting concerns, explaining the strong intuitions about reasonable bets as coming from ideas that are more related to guidance, access, and internalism. That is, one can deny that the degree of belief in a proposition that would be the best (epistemically) in a certain situation need correspond

to the odds at which it would be rational to bet on that proposition, at least if we're thinking of bets in a first-personal way. There is, however, a very important caveat to this. I am here rejecting the close connection between a reasonable bet on *P* and a reasonable degree of belief in *P*. However, one could still think that if a set of bets collectively result in a *certain* loss, then this *does* show something of interest. Though according to the line I am pushing it is illegitimate to criticize the agent for any particular degree of belief in a proposition based on betting considerations, it may still be appropriate to criticize an agent for having a set of credences such that betting according to them results in a certain loss. This is because this latter situation shows us something about the consistency of a set of degrees of belief. Thus, one can accept that Dutch Book considerations do tell us something interesting even while objecting that the simple question, 'Would you bet on that at those odds?' is not appropriate in the context of this project.[29]

3.6.2.3 The sq-*COND* Framework

A different way of pressing this worry doesn't go via betting considerations. Instead, we simply ask: if something is indicated by a process with less than maximal reliability, then why *should* someone respond by giving it credence 1? RAE says that agents are sometimes epistemically required to give credence 1 to propositions that are indicated by less-than-perfectly reliable belief-forming processes. At first glance, this is clearly wrong.

It is important to note that this problem, if it is a problem, affects *many* different views, and not just RAE. For every belief is formed by some process, and it is very

[29] For more detailed discussion of betting considerations and Dutch Book arguments, see Chapter 6.

plausible that *all* of them are less than fully reliable. If so, then *any* account of evidence that has the consequence that we do have some evidence is committed to saying that less-than-perfectly reliable beliefs are evidence. So long as we work within a framework where evidence gets credence 1, we have the result that less than perfectly reliable beliefs get credence 1. This is true, for instance, even for a paradigmatically internalist view like MSE. RAE perhaps brings this more obviously to the forefront, but it is present in *many* accounts of what it is to have evidence.

To make this point more forceful, consider several accounts of evidence different than RAE. Consider first the deflationary account of evidence, which says that an agent's evidence consists of all and only the propositions to which the agent gives credence 1. Clearly, it is possible for an agent to give credence 1 to proposition in a less-than-perfectly reliable way. So, according to the deflationary account of evidence, less-than-perfectly reliable beliefs get credence 1.

Consider an internalist account of evidence according to which the proposition that *P* is evidence just in case it appears to the agent as if *P*. Since it can appear to me as if *P*, even if *P* is false, such an account says that it is possible that *P* is evidence even though it is less-than-perfectly reliably indicated to me.

Consider a variant internalist account of evidence according to which the proposition "It appears as if *P*" is evidence just in case it appears to the agent as if *P*. Now, if I have a process that produces full credence in the proposition "It appears as if *P*" iff it appears to me as if *P*, then this would be a perfectly reliable process. However, we don't have such processes. There is not an infallible link between our beliefs about

how things appear and how things actually appear.[30] So, imagine a time when it appears to me as if *P*, and yet the link between the appearances and the belief about the appearances is not there. This account of evidence says that whenever it appears to me as if *P*, then I should give credence 1 to the proposition that it appears to me as if *P*. So, this time I should give credence 1 to the proposition that it appears to me as if *P*. But this is a time when the link between the belief and the appearance is absent. So, this account says that less-than-perfectly reliable processes can result in credence 1.

Given that this situation arises for so many accounts of evidence, we do not have here a problem specific to RAE. But it is too dismissive to say that there is *no* problem. What we might have is a much broader objection to any account of evidence according to which agents have contingent propositions as evidence and yet evidence receives credence 1. So, though this worry doesn't put pressure on RAE *per se*, it does put pressure on the idea that evidence should get credence 1. This brings up an important motivational issue for anyone defending RAE: why adopt a framework where evidence gets credence 1? This is a good question. In what follows I'll give three potential answers, and show why they all seem to fail. Then I'll suggest a fourth answer, that though not terribly satisfying, may be a good answer nevertheless.

3.6.3 Why Adopt a Framework Where Evidence Gets Credence 1?

Answer 1: *It is intuitively obvious that our evidence comes in the form of certainties, and not in degrees.*

There is something to this answer to the question. It does, I think, feel as if our evidence in a given situation comes in the form of certainties. Let's say I'm sailing on a

[30] This claim is defended in more detail in Chapter 4.

dark night and trying to figure out where we are on the charts. I look off the starboard side and see a lighthouse's signal. I then use this information—that there's a lighthouse to our starboard side—as evidence. And so long as it is part of my evidence that there is a lighthouse to our starboard side, it seems as if it is certain evidence. This is not to say, of course, that the situation might not be different. Consider a second scenario in which I'm only 90% confident that there is a lighthouse to our starboard. In such a situation it is natural to say that my evidence is something else, something weaker—perhaps that it appears as if there is a lighthouse to our starboard side. But this *is* certain evidence.

So, I think it is natural to think that our evidence comes in the form of certainties. But I don't want to put much weight on that, for it seems a very unstable intuition. When it comes to considering the possibility of uncertain evidence, it is very easy to get confused. For one person might assert that one's evidence comes in the form of certainties, and represent the second lighthouse scenario as:

Evidence: *Appears Lighthouse*

cr(*Lighthouse*|*Appears Lighthouse*) = 0.9

But someone who thinks our evidence is graded might describe the situation as simply one where the agent has as evidence that there is a lighthouse to degree 0.9. We could represent this as:

Evidence: <*Lighthouse*, 0.9>

I do not think that our intuitions really favor one of these representations over the other, so I do not think we can be sure that it is intuitively obvious that our evidence comes in the form of certainties.

But there is a second, more important problem with this answer. The account of evidence that I am trying to provide is a normative account of evidence. It is trying to tell us what our evidence *really* is in a given situation. So let's grant that it is not only intuitively obvious that our evidence comes in certainties, but that it is true that our evidence comes in the form of certainties. This, on its own, doesn't show anything. For it might be true that we all affirm the consequent. This would not mean that the correct normative account of deductive inference would need to be one that allows us to affirm the consequent. So, this answer to the question fails.

Answer 2: *It is easier for agents to reason if evidence gets credence 1, rather than if agents receive graded evidence.*

There is something to this answer, too. It is certainly easier for me to reason in a given situation if I take my evidence as receiving credence 1. This drastically limits the complexity of the computations that need to be made in Bayesian reasoning. However, this answer also fails to be a convincing answer to the question. For consider our discussion of the OIC arguments in Chapter 1. There it was made clear that we are pursuing an evaluative epistemic account, that eschews issues about computational complexity. So ease-of-use considerations are not going to motivate working within the evidence-gets-credence-1 framework for the project we are engaged in.

Answer 3: *We can mimic whatever the graded-evidence accounts can do within the framework where evidence gets credence 1, just by identifying special evidence propositions.*

Again, there is something to this answer. Let '*Green*' = the cloth is green, '*Red*' = the cloth is red, and '*Blue*' = the cloth is blue. Imagine that we want to model a situation where the agent looks at the cloth under candlelight, and his evidence is:

<*Green*, 0.4>, <*Red*, 0.2>, <*Blue*, 0.4>

We can do this, while only countenancing evidence that gets full credence by thinking of the agent's evidence as the proposition, *E*, such that:

cr(*Green*|*E*) = 0.4, and cr(*Red*|*E*) = 0.2, and cr(*Blue*|*E*) = 0.4

Using this trick, we can mimic graded evidence with evidence that only ever gets credence 1.[31]

This, however, is an answer that the defender of RAE must treat with skepticism. For according to RAE, a proposition is evidence only if there is a reliable belief-forming process that indicates that proposition. But what is this proposition *E* and why do we have any reason to think that there is a reliable mechanism that indicates *E*? These are difficult questions.

Timothy Williamson ([2000]) has given a response to questions much like these. He has noted that there need be nothing exotic about propositions like *E*, so long as the propositions in a credence function can use demonstratives. *E*, for instance, might be something like: "It looks like *that*," where '*that*' refers to the qualities of the agent's visual perception. There is good reason to think that there might be a reliable mechanism indicating such an *E* to the agent.

However, there are difficulties with this approach. First, it is not clear how to correctly define a credence function over propositions that include demonstratives. For

[31] Jeffrey ([1965]) notes this in chapter 11.

if credence is ever given to a proposition, then the credence function is always defined over that proposition. But does it make sense to ask about my credence now that it looks like *that*, where '*that*' refers to the qualities of some visual perception I will have in the future?[32] If we can't make sense of this, then Williamson's response faces a technical difficulty.

Bracketing this concern, however, it is not clear that this proposal will really get around all the worries. For above we noted that there is reason to think that *no* contingent propositions should get full credence. Though Williamson's move may be plausible in the cloth case above, it isn't clear that it will always work. For what happens when we think that the proposition with the demonstrative (*E* in the case above) should graded? To make a move here analogous to Williamson's, the evidence propositions we would have to countenance really do look mysterious. So, it seems, the problem simply reasserts itself. Thus, this third answer is inadequate.

I have looked at three possible answers to the question posed and found them all to fail. So, what *is* the motivation for going with a framework where evidence gets credence 1?

Answer 4: *This framework is a good starting point. It would be good if it were expanded to handle graded evidence. But such an expansion faces grave difficulties.*

Unlike the three answers above, there isn't much to this answer. It acknowledges that there *is* something odd about the evidence-gets-credence-1 framework when

[32] Note, that the problem I am highlighting would not be solved by an account of *de se* credences (as discussed in, for example, Meacham ([*forthcoming*]) and Kim ([2009]). What I am drawing attention to has to do with assigning a value to a proposition that doesn't seem to have any content until the experience referred to is present. This is separate from the issue of self-locating belief.

combined with the clear truth that our epistemic faculties are fallible. However, it defends working within that framework for largely theoretical reasons.

This answer, however uninspiring, is made more appealing by noting the difficulties facing any attempt to give an account of graded evidence. In what follows I will explain how one might construct an account of graded evidence along reliabilist lines, and then point out the difficulties that such an account will face. I will then generalize these difficulties to show that it is a feature of graded evidence, not the reliabilist approach. This will motivate the focus on accounts of evidence that work within the framework where evidence gets credence 1.

3.6.4 On Graded Evidence

Reliability information about processes of belief-formation comes in probabilistic form. When we say that a process is 0.9 reliable, we mean that 90% of the time the process leads to the belief that *P*, *P* is true. It is natural to use this reliability information to give us an account of graded evidence. A first idea is that if a process indicating *P* is *n* reliable, then the agent's evidence is <*P*, *n*>. This simple idea quickly runs into problems. In particular, since *every* proposition indicated to an agent is indicated with *some* level of reliability, every proposition in the agent's credence function will be paired with a reliability number, thus making all the agent's credences evidence. This is unacceptable.

But there is a more complicated reliabilist idea that works better. The idea is similar in many ways to RAE, but with important differences. First, instead of considering belief-forming processes, we will now consider credence-*n* producing processes. Second, we will introduce the notion of calibration. Imagine that a weather

forecaster forecasts the chance of rain over many different days. We'd like to know something about how good a forecaster he is. Here is one way to evaluate him. Take all the days when the forecasted chance of rain was *n*%. If the proportion of those days that were rainy is *n*, then he is perfectly calibrated. Of course, he might not be *perfectly* calibrated, but still be better or worse calibrated. As the proportion of those days that were rainy moves further away from *n*, the weather forecaster is less calibrated; as the proportion moves closer to *n*, the weather forecaster is more calibrated. Now, there are various non-equivalent ways to define this notion of calibration. But the basic purpose of these definitions is the same: to formalize a way to rate the accuracy of probabilistic judgments. For our purposes, we won't need the technical details, only the rudimentary idea that a probabilistic judgment can be more or less calibrated, and that the level of calibration of one judgment can be meaningfully compared with the level of calibration of another judgment.

Given this, we can now state a reliabilist account of graded evidence:

RAGE: $\langle E, n \rangle$ is evidence iff there is an available process calibrated at level *c* or above that results in $\mathrm{cr}(E) = n$.[33]

Now, we could state RAGE in such a way that perfect calibration is required (so that $c = 1$), but we need not do this. Note that if we do not do this, then RAGE has RAE as a special case. If we set $n = 1$ in RAGE then we have:

E is evidence iff there is an available process calibrated at level *c* or above that results in $\mathrm{cr}(E) = 1$.

[33] I have left out time indices since they are not our main concern here.

But what is it for a process resulting in cr(E) = 1 to be well-calibrated? It is for E to nearly always be true when the process results in cr(E) = 1. But this is just to say that the process must be reliable. So, RAGE can be seen as a *generalization* of RAE.

Now, how does RAGE work? Well let's say that I look out in the garden and see a finch. Let's say that I'm pretty good at this, say, 0.99-reliable at identifying finches. RAE says that F: "there is a finch", is thus a member of my evidence set. But RAGE need not say this. Suppose that I have a finch-identifying process that results in credence of 0.99 that there is a finch. Further, let's imagine that the proportion of times when there is a finch out of all those times when this process works is 99/100. Then, RAGE will say that <F, 0.99> is evidence for me. This is different than RAE which says that F is in my evidence set (and so assigned credence 1). This gives one a feel for how RAGE works.

Now, there are some minor technical issues to work out with RAGE. For instance, what do we say about cases where there are two available processes, both well-calibrated: one that leads to cr(F) = 1, and one that leads to cr(F) = 0.99? Further, one might be concerned that RAGE is based on the idea that we have credence-n producing processes. But what if we don't have such processes? While it is uncontroversial that we have *belief-forming* processes (as RAE requires), it is much less certain that we have credence-n producing processes, especially to the degree of precision required by RAGE. These are tricky questions. However, I will not attempt to answer them, because I think any normative account of graded evidence will run into serious problems more general than these. This will make good on my claim above that any account of graded evidence faces grave difficulties.

The problem can best be illustrated by thinking about different kinds of accounts of graded evidence that one might have.

3.6.4.1 Type 1 Theories

Type 1 theories are the most natural starting place for an account of graded evidence. According to such theories, there is a correspondence between features of the situation at a time and proposition-degree pairs that represent the agent's evidence. Further, there is a straightforward relationship between proposition-degree pairs representing evidence, and the effect that this evidence should have on the agent's doxastic state. According to Type 1 theories, if $\langle P, n \rangle$ is evidence at t, then this requires that $\mathrm{cr}_t(P) = n$. So, according to Type 1 theories, the presence of some feature F of the situation at t determines that some proposition E_F is evidence to degree n such that for a large class of credence functions, if F is present, then $\mathrm{cr}_t(E_F) = n$. The idea behind such a theory is that we can compositionally construct the agent's evidence at a time from different features of the agent's situation at that time. Since feature F determines the evidence for a wide class of credence functions, we can do this in a general, and thus informative way.

There is, however, a problem with Type 1 theories. Consider a situation where F and G are both present, and where the theory tells us that F results in $\langle E_F, 0.9 \rangle$ and G results in $\langle E_G, 0.8 \rangle$. Now, it is perfectly possible for this to effect a change in the agent's credence function such that $\mathrm{cr}(E_F) = 0.9$ and $\mathrm{cr}(E_G) = 0.8$. The problem is that this underdetermines the agent's update. There are many ways of changing credence so that $\mathrm{cr}(E_F) = 0.9$ and $\mathrm{cr}(E_G) = 0.8$. To uniquely determine how the agent is to update, we would need an evidence *partition*, of the form:

$$<E_F \wedge E_G, a>; < E_F \wedge \neg E_G, b>; <\neg E_F \wedge E_G, c>; <\neg E_F \wedge \neg E_G, d>$$

with the appropriate values filled in for *a, b, c,* and *d*. But a Type 1 theory need not give us this, so it is incomplete.[34]

3.6.4.2 Type 2 Theories

A natural response to this problem is to pursue a slightly different sort of theory. Instead of considering partial features of a situation, we now consider the total feature of a situation at a time, *TF*, which determines the evidence *partition* at that time. The presence of feature *TF* of the situation at a time determines that some evidence partition $\{E_i\}_{TF}$ is evidence with values n_i such that when *TF* is present $\text{cr}(E_i) = n_i$.[35]

This solves the problem above, because we are never in a situation where the partition isn't specified. However, we still face two problems, one technical, and one theoretical. The theoretical problem is that if the relevant feature *TF* is *too* specific, then our theory will be simply an *ad hoc* categorization of every possible epistemic situation and the evidence partition for that situation. This loss of generality makes an account less a theory and more a stipulation. On the other hand, if we let the relevant feature *TF* be too general, then we will run into situations where there is more than one such "total feature" present, and we will not have avoided the problem with Type 1 theories. This inability to compositionally construct what an agent's evidence is from various features of the situation is a major drawback for Type 2 theories.

---

[34] Something along these lines is suggested, very briefly, in Levi's ([1967]) review of Jeffrey's *The Logic of Decision*.

[35] This is naturally combined with Jeffrey Conditionalization: $\text{cr}_t(\bullet) = \sum_i \text{cr}(\bullet|E_i) \times \text{cr}(E_i)$
where the E_i together form a partition over propositions representing the agent's evidence.

The technical problem is the well-known problem of non-commutativity noted first in Field ([1979]).[36] The problem can be best illustrated by an example. In a wide range of cases, it seems that the order in which an agent receives information should be irrelevant to the final distribution of credences. Consider a case where I first hear rain hit the window, and then see the rain falling outside, compared to a case where I first see the rain falling outside and then hear the rain hit the window. It seems that my final credences should be the same in each situation. But a theory of Type 2 is not, in general going to deliver this result. Let TF_h be the total feature of the situation when I hear the rain and TF_s be the total feature of the situation when I see the rain. For definiteness, let's assume that the evidence partition given by TF_h implies <*rain,* 0.8> and that the evidence partition given by TF_s implies <*rain,* 0.9>. Given this, my final credence in *rain* in the scenario where I first hear the rain must be 0.9, and my final credence in *rain* in the scenario where I first see the rain must be 0.8. So, we have a failure of commutativity: simply because we changed the order of the learning episodes, we changed the final distribution of credences.

3.6.4.3 Type 3 Theories

A well-known fix to this problem with commutativity is to rethink what an account of evidence should be doing. According to Type 1 and 2 theories, an account of evidence takes us from some feature of the situation, to the *final evidential effect* that that feature should have on a credence function. The evidence-acquisition episode—or *learning episode*—is completely described by the final distribution of credences that episode of learning results in. According to Type 3 theories, on the other hand, some total feature

[36] See also Doring ([1999]), Lange ([2000]), Wagner ([2002], [*forthcoming*], and Weisberg ([2009]).

of the situation at a time determines the *evidential impact* that that feature should have on a credence function, though it does not tell us the final result of this impact. The learning episode is completely described by the *evidential impact* that feature *TF* makes, but not the final distribution of credences. A particularly popular way of implementing such a strategy is to think of the feature *TF* as determining the evidence partition and the Bayes' Factor for that partition.[37] These two things together encode the evidential impact of *TF*. Consider a simple situation where the evidence partition is $\{E, \neg E\}$, and the evidence is acquired between time 1 and time 2. The Bayes' Factor for this is given by:

$$\beta_{cr2,cr1}(E{:}\neg E) = \frac{cr_2(E)/cr_2(\neg E)}{cr_1(E)/cr_1(\neg E)}$$

According to this view, *TF* gives us the partition and the values of the Bayes' factor, then we use this together with the agent's previous credence function (cr_1) to determine the final evidential effect that the feature has on the agent's belief state.

Such an account of evidence will allow evidence to commute in the sense that updating some credence function first as a result of learning episode E and then F leads to the same distribution of credences as updating that function first as a result of learning episode F and then E. In fact, Wagner ([2002]) has recently proven that adopting a theory of Type 3 with Bayes' Factors encoding the evidential impact is the *only* way to escape commutativity worries.[38]

However, Type 3 theories also face a significant problem, first noticed by Garber ([1980]). The problem is that Type 3 theories allow an agent to accumulate near-

[37] Field ([1979]) and Wagner ([2002]) each propose this.
[38] This claim actually requires several plausible assumptions. See Wagner ([2002]) for details, or Weisberg ([2009]) for a more accessible discussion of the proof.

certainty in a certain proposition in rather impoverished evidential situations through simply re-inspecting particular scenes. Wagner summarizes the problem as follows:

> Imagine that you glance briefly in dim light at an object known to be blue or green, resulting in your becoming slightly more confident than you were before that the object is blue. Repeated glances producing the identical sensory stimulus will then result in your approaching certainty that the object is blue. (Wagner [2002], p. 276)

That is, according to Type 3 theories, each time you glance at the object, the hypothesis that it is blue gets a little probabilistic bump. Repeating this process by glancing repeatedly at the object gives further probabilistic bumps to the hypothesis, until we approach certainty that the object is blue.

Now, for Type 3 theories to have this consequence, it must be the case that in the repeated glances, the relevant total feature *TF* the same. One might respond to this problem by claiming that in each glance, the relevant total feature is different. If that's the case, then Type 3 theories need not say in this case that we can build up certainties through repeated glances. In the scenario I am considering, however, I stipulate that the environment and the glance are exactly the same from glance to glance. Thus, the difference in total feature, *TF*, from glance to glance must either lie in the agent's credence function itself, or in some historical fact about the situation. If the difference is the historical fact (e.g., that the agent *already* glanced at the scene), then we are not giving a Type 3 theory, which attempts to give an account of evidence at a time based on features present at that time. This is problematic because once we allow historical facts like this to be relevant to the agent's evidence then we run into the theoretical

problem discussed with Type 2 theories. We will end up giving an account of evidence where, given the total life history of an agent, we state what the agent's evidence is at each time. But there is no reason to think, that such an account will look anything like a theory, rather than a stipulation.

Perhaps, then, the difference in the total feature concerns the agent's credence function itself. It is, however, not clear how this would go. The most plausible way to think how the agent's credence function is relevant in this situation is to think that in subsequent glances, the agent has the information that he has already glanced at the same scene. However, this is just a bit more evidence that the agent has. So appeal to such features to understand the agent's evidence presupposes that we already understand the agent's evidence. Thus, Type 3 theories are inadequate.

3.6.4.4 Summing Up Graded Evidence

What type of theory could get around these problems? There seem to be two options, neither of which are promising if our goal is to give a *normative theory* of evidence. One option is to change our perspective about evidence and credal change. Instead of looking to an account of evidence to understand reasonable credal change, we look at reasonable credal change to understand evidence. By looking at how an agent's credences ought to evolve in different situations, we pin down what the agent's evidence must have been in those situations. The idea here is not simply the familiar one where we use intuitions about reasonable credal change to home in on an adequate account of evidence. According to this approach, there are no theoretical generalizations made about evidence in various situations. Instead, our convictions about reasonable credal change directly determine what the agent's evidence is: the agent's evidence at a

time is whatever it is that updating on would have resulted in the rational credal changes at that time. So, if in repeated glances at the object, we think it obvious that an agent's credence in it being blue ought not change after the first glance, *this* fact itself determines what the agent's evidence is in that particular situation. But there is no explanation offered for *why* this is. It simply follows from our intuitions about rational credal change that things must be like this.[39] The problem with this is twofold. First, though our intuitions about reasonable belief change in some situations are firm, in many other situations they are inchoate, so this approach offers little guidance to evaluators. Second, this is hardly a *theory* of evidence. Instead it is a stipulative endeavor guaranteed by fiat to get intuitions right.

The other option is closely related, and is the idea that we simply drop the search for a normative account of evidence altogether. Though it is hard to tell for sure, this idea appears to be Wagner's ([2002]) approach. His guiding idea is that sameness of learning episodes is reflected in identical Bayes' Factors for an evidence partition. So, whenever there is the same Bayes' Factor for an evidence partition we can say that the same evidence was acquired. But we can't appeal to general features of the situation to figure out what evidence was acquired and so if there *should* be the same Bayes' Factor and partition. As he writes, "It is important to note here that individuals can have the same *isolated* sensory, or other, experience without undergoing what I have chosen to call identical *new learning*." ([*forthcoming*], footnote 8). There is nothing we can appeal to to figure out when such *new learning* has occurred outside of intuitions about rational credal change and perhaps particular intuitions about what evidence an agent has. If we

[39] This kind of approach appears to be in agreement with the approach advocated in Neta ([2008]).

take this route we again appear to be rejecting the idea that we can give a normative theory of evidence.

I've looked at different general types of theories that might be offered as accounts of graded evidence. Type 1 theories ran into the problem with combining evidence at a time. Type 2 theories ran into theoretical problems concerning how to specify the evidence at a time in a way that is theoretically grounded. Type 2 theories also ran into a problem with combining evidence, this time evidence gathered at different times. Type 3 theories resolve this latter problem, but run into the problem discussed by Garber ([1980]). The alternative is to reject the idea that we could have an interesting normative theory of evidence. This certainly is an alternative, but it is not a welcome one. For it seems that we should be able to come up with such a theory. At any rate, I am working under the assumption that such a theory is possible. If we are working under this assumption, then I have provided a reason for sticking with the sq-*COND* framework. For if my arguments here are correct, then there are problems with getting a normative theory of graded evidence that have nothing to do with the specific features of the theories proposed. In light of this, I think that Answer 4 (Section 3.6.3) looks better than it might have at first glance.

One might be curious why the problems for graded evidence presented here do not arise in the sq-*COND* framework. Type 1 theories struggle to combine different pieces of evidence. In particular, we noted that if feature *F* mandates $<E_F, n>$ as evidence and feature *G* mandates $<E_G, m>$ as evidence, we don't have all the information that we need. What we need is the values *a, b, c,* and *d* for:

$<E_F \wedge E_G, a>$; $< E_F \wedge \neg E_G, b>$; $<\neg E_F \wedge E_G, c>$; $<\neg E_F \wedge \neg E_G, d>$

But we don't run into this problem when evidence gets credence 1 since in this case $n = m = 1$, and when this is the case, *a, b, c,* and *d* are uniquely determined: $a = 1$ and $b = c = d = 0$. Further, when it comes to issues of commutativity, order doesn't matter so long as evidence propositions receive credence 1. So, according to a framework where evidence receives credence 1, we don't need Bayes' Factors.

One final word is necessary before leaving the topic of graded evidence. I don't take what I have said here to be a criticism of much of the interesting work done on graded evidence, which often concerns how to incorporate such evidence into an existing probability function. What I think this criticism *does* show is that it is unlikely that we will get a nice normative theory of when an agent has graded evidence.[40] But my claim is not that such a theory would be unnecessary or unwanted. Such a theory would be very good to have. It just may not be one that we can get. In light of this, I focus on giving a theory within the sq-*COND* framework.

3.7 Conclusion

In this chapter I have proposed RAE, which is an account of evidence based on reliability considerations.. The bulk of the chapter has been clarificatory, explaining the account, and responding to objections based on such clarification.

I first noted how accounts of evidence are under pressure to satisfy two desiderata: the idea that evidence connects us to the world, and the idea that evidence provides guidance. I explained that a reliabilist account emphasizes the former of these, sometimes at the expense of the latter. I explained why some such trade-off is to be expected. Next I stated RAE, and clarified how to understand some of its key terms,

included ‘belief-forming process’ and ‘availability’. After this I discussed the generality problem, which afflicts all forms of reliabilism, and offered what I see as a promising response to this problem.

I concluded with a long section that discussed the two claims made by RAE:

(1) less than fully reliable processes can yield propositions as evidence, and

(2) the reliability of the threshold should be a variable threshold depending on the context of evaluation.

Discussion of these two claims involved us in both small issues as well as very large ones. I offered some justification for proceeding in the sq-*COND* framework, according to which evidence gets credence 1. Using this, I explained how (1) and (2) are defensible.

[40] This general point is in agreement with the line taken in Christensen ([1992]).

CHAPTER 4

DEFENDING AND MODIFYING RAE

4.1 Introduction

In the last chapter RAE was stated and clarified. This chapter is devoted to the evaluation of RAE.

A large part of this evaluation will concern the presentation of objections and responses. The objections form two somewhat natural sets. The first set of objections include objections directed at any theory that is a type of Evidential Externalism (EE). It is possible that the only available processes of belief formation that are reliable enough to yield evidence according to RAE are all processes that treat internal twins the same. If this were the case, RAE would not necessarily be a type of Evidential Externalism. However, this certainly need not be the case. Thus, objections to EE seem to be objections to RAE. The second set of objections include objections directed specifically at RAE. I will first consider the more general objections (in the first group), and then the more specific objections (in the second group) after this.

Though I will offer responses to all these objections, some might still doubt the tenability of an account like RAE. Thus, though I defend RAE, at the end of this chapter I'll investigate a reliabilist *constraint* on evidence (RCE), where reliability is only a necessary condition on an agent having evidence. Since many of the objections to RAE aim at discrediting reliability as *sufficient* for having evidence, moving to RCE would allow one to sidestep these. Accordingly, as the objections to RAE are encountered, I'll note how and if a move to RCE would block the objection.

I will close the chapter by discussing attractive features of RAE, and to what extent the weaker RCE possesses these features.

4.2 Objections to Evidential Externalism

4.2.1 Silins's Good/Bad Case

4.2.1.1 Silins's Argument

Silins ([2005]) offers a general argument against any form of EE, in favor of EI. Since RAE is a form of Evidential Externalism, it is a target of Silins's argument. Silins's main argument against EE turns on showing that it is a consequence of EE that a radically deceived subject can be epistemically better off than a non-deceived internal twin of that subject. This is thought to be an intolerable consequence. Since it is entailed by EE, EE must be rejected.

Silins's argument takes the following form:

(1) If EE is true, then a subject in the bad case is sometimes epistemically better off than his internal twin in the good case.

(2) A subject in the bad case is never epistemically better off than his internal twin in the good case.

(C) Thus, EE is false.

To understand this, we need to know something about what it is to be in a bad case and what it is to be in a good case. The basic idea is as follows: If A has an internal twin B, and B is somehow deceived, then A is in the good case relative to B who is in the bad case. For example, if things are as we think they are, then you are in the good case, relative to a brain-in-a-vat that is internally identical to you.

Silins presents the following scenario to support the premises of this argument. Gary is in the good case. Let *B* = "I had a banana for breakfast yesterday." Imagine that *B* is true when considered by Gary. Further, assume that EE is true and that the right external relation obtains between Gary and the fact that *B* so that *B* is a member of Gary's evidence set. Perhaps Gary *knows* that *B*. Or perhaps Gary remembers that *B*, and memory is sufficiently reliable to render *B* a member of his evidence set. Let *SB* = "I seem to remember that I had a banana yesterday." We assume that *SB* is true when considered by Gary, and that the right relation obtains between Gary and the fact that *SB* so that *SB* is a member of Gary's evidence set. So, both *B* and *SB* are members of Gary's evidence set according to EE. However, realizing his memory can sometimes be mistaken Gary is such that $cr_t(B) = 0.9$, while $cr_t(SB) = 1$.

Now consider Barry, who is unfortunately nothing more than an envatted brain, and happens to be Gary's internal twin. Barry is in the bad case. He didn't actually have a banana yesterday, so *B* is false when considered by Barry. Further, we stipulate that the right external relation does not obtain between Barry and the fact that *B*. Accordingly, *B* is not a member of Barry's evidence set. Nevertheless, as he is an internal twin of Gary, Barry does seem to remember having a banana. Since all the internal facts about Gary are true of Barry, *SB* is a member of Barry's evidence set. Further, Barry is such that $cr_t(B) = 0.9$ and $cr_t(SB) = 1$, again because they are internal twins.[1]

[1] Note that *B* and *SB* are both propositions which feature the indexical expression "I". By characterizing Barry and Gary as having beliefs in propositions like *B* and *SB*, Barry and Gary seem like internal twins. Of course, there is a sense in which when Barry believes *B* he is believing something very different than when Gary believes *B*. I note this complication only to set it aside, as it doesn't seem to play any role in the discussion. For our purposes, if Gary and Barry both believe *B* and *SB* to the same degree then they have identical doxastic states.

Now, let us stipulate that that Barry and Gary are both such that $cr_0(B|SB) = 0.9$. Then we get the following result: Barry has the credences that result from conditionalizing on Barry's evidence, but Gary doesn't have the credences that result from conditoinalizing on Gary's evidence. Gary's $cr_t(B) = 0.9$, even though *B* is evidence for Gary. Thus, Barry is doing epistemically better than Gary is in this scenario, even though Barry is in the bad case, and Gary is in the good case. This supports premise (1).

The support for premise (2) is more sparse. Silins claims that this kind of situation violates a plausible principle:

> **Bad Case Principle (BCP)**: Necessarily, if B is in the bad case and A is an internal twin of B in the good case, B is not more justified in believing *P* than A. ([2005], p. 389)

One might take the story about Gary and Barry to provide some support for BCP. Perhaps it is supposed to follow clearly from the story that Gary is *not* worse off than Barry, and so provide some inductive support for BCP. Other than this, it is simply claimed that BCP is a plausible thesis. BCP, however, can be challenged.

4.2.1.2 Against the Bad Case Principle

I think the plausibility of Silins's argument comes largely from the thought that Gary really *isn't* doing anything epistemically incorrect in the situation described. And if he isn't doing anything epistemically incorrect, then it's hard to see how Gary is anything but fully justified in his belief about having a banana. And if that's the case then it's *impossible* that a brain-in-a-vat could be doing *better*. If Gary's doing as good as can be

done, then *no one* could be doing better. We then generalize this intuition, and arrive at BCP.

So why might one think that Gary really isn't doing anything epistemically incorrect in the scenario above? This verdict is delivered, I propose, as a result of reflection about reasonable doubts. Gary, one might think, reasonably doubts his memory in the scenario Silins describes. Thus, he is doing something epistemically *appropriate* by refusing to fully believe that he had a banana yesterday morning. However, I'll argue that there are cases structurally identical to the one that Gary finds himself in, where such doubts do not look so appropriate. The existence of such a case shows it is possible for someone in a good case to be doing epistemically worse than his envatted twin in the bad case. Note that we should expect such a case so long as it is reasonable to maintain that one can be epistemically deficient for a failure to respond to one's evidence. Such a case will show the Bad Case Principle to be false.

Here is one such case. Let R = "there is a red stop sign in front of me." Assume that there is a red stop sign in front of Gina. She's looking right at it, and the lighting is ideal. Gina has never been wrong about the presence or absence of stop signs in such conditions. As one might expect, the adherent of EE holds that in this situation the right external relation obtains between Gina and the fact that R so that R is a member of Gina's evidence set. Let SR = "I seem to see a red stop sign." SR is also true of Gina and the right external relation obtains between Gina and the fact that SR so that SR is a member of Gina's evidence set. So, both R and SR are members of Gina's evidence set. However, Gina is such that $cr(R) = 0.9$ and $cr(SR) = 1$.

Beth is an envatted brain in the bad case. Given this, there is no red stop sign in front of Beth and the right external relation does not obtain between Beth and the fact that *R*. Accordingly, *R* is not a member of Beth's evidence set. Nevertheless, as she is an internal duplicate of Gina, Beth does seem to see a stop sign, and the right external relation obtains so that *SR* is a member of Beth's evidence set. Again, since they are internal twins, Beth is such that cr(*R*) = 0.9 and cr(*SR*) = 1.

Now, let's stipulate that before looking at the stop sign, Beth and Gina were both such that cr(*R*|*SR*) = 0.9. But then we get the following result: Beth has the credences that result from conditionalizing on her evidence, but Gina doesn't. Gina's cr(*R*) = 0.9, even though *R* is evidence for Gina. Thus, Beth is doing better than Gina in this scenario, even though Beth is in the bad case, and Gina is in the good case.

I submit that in this case it is appropriate to say that Gina is doing epistemically worse than Beth. The reason for this is that it is not at all obvious in this case that Gina should entertain doubts with respect to *R*. Further, there's a clear sense in which Gina really *is* doing worse here. There is a red stop sign directly in front of her, the environment is ideal, and yet Gina doesn't treat this as evidence. There is a clear sense in which Gina would be doing better if she believed *R* to degree 1. Further, *in this same sense*, Beth is fully justified in believing *R* to degree 0.9. Beth has only *SR* as evidence, and so conditionalizing on this results in cr(*R*) = 0.9. Thus, if Gina were such that cr(*R*) = 1, this would be just as justified as Beth's cr(*R*) = 0.9. For each of them, this is the credence that results from conditionalizing on their evidence. Now, since Gina and Beth are internal twins, they cannot give different values to cr(*R*). Altering Gina's cr(*R*) from 1 to 0.9, makes Gina and Beth internal twins. But surely altering Gina's belief state in

this way doesn't change Beth's degree of justification. So, in doing this, Gina's degree of justification must change. But it can't change for the better, for she was fully justified to begin with, so it must change for the worse. So, in the scenario, it is appropriate to say that Gina is less justified than her brain-in-a-vat twin, Beth. The Bad Case Principle says this can't happen, so it is false.

One may be unconvinced by this example. In particular, one might insist that Gina really wouldn't be doing better to have full credence in *R*. Intuitively, you might say, Gina doesn't have *R* as evidence as I claim that she does. That evidence is not available to her, and she simply entertains reasonable doubts about her perceptual abilities.

If this is one's reaction to the example, then there is an alternative response to Silins's argument. This second response takes the form of a *tu quoque*. Essentially, the reasoning goes, if you think that Silins's argument is successful and if you insist that it remains so in the stop sign example, then there are structurally similar arguments that push you towards more and more internal accounts of evidence, which are untenable.

4.2.1.3 Tu Quoque

Suppose I define two agents to be internal duplicates just in case their non-factive *doxastic* states are identical. Call such twins d-internal twins. This is different than MSE (and Silins's preferred account of evidence), which says that two agents are internal duplicates just in case they share all non-factive *mental* states.[2] Call such twins m-internal twins. Silins is an Evidential Internalist. He holds that it is not possible for there to be evidential differences between m-internal twins. Imagine that I, too, am an

Evidential Internalist, but of a somewhat different stripe. I hold that it is not possible for there to be evidential differences between d-internal twins. I ascribe to Doxastic Evidential Internalism (dEI) whereas Silins ascribes to Mental Evidential Internalism (mEI).

Now, mEI and dEI appear to be different theses. They appear to offer accounts of evidence based on two different supervenience bases. But, if we follow Silins, then mEI and dEI are stated as follows:

mEI: Necessarily, if A and B are m-internal twins, then A and B have the same evidence.

dEI: Necessarily, if A and B are d-internal twins, then A and B have the same evidence.

If A and B are m-internal twins, then they are d-internal twins. So if mEI says that subjects have the same evidence, so does dEI. But there are situations where A and B are d-internal twins but not m-internal twins. dEI must say that in such situations A and B have the same evidence. Officially, mEI takes no stand in such cases, since these are simply situations where the antecedent of mEI is false. However, since an account of evidence based on m-internalism is supposed to be an interestingly different than an account based on d-internalism, this difference should sometimes be manifested. If it *weren't*, then mEI would just be a misleading way of expressing dEI. So, there should be some cases in which dEI says that two subjects have the same evidence, and yet mEI disagrees. To bring this out, I'll state mEI and dEI in a stronger way:

mEI*: Necessarily, A and B are m-internal twins iff A and B have the same evidence.

---

[2] See Chapter 3, section 3.1.

dEI*: Necessarily, A and B are d-internal twins iff A and B have the same evidence.

I realize that mEI* and dEI* may be untenable. For example, a sensible version of mEI is *not* committed to the claim that *any* m-internal difference entails an evidential difference. For instance, *P* might be evidence for subject A because A seems to remember that *P*, whereas *P* might be evidence for subject B because B seems to experience *P*. Perhaps there can be m-internal differences without a difference in evidence. Nevertheless, if some form of mEI is going to be interestingly different than some form of dEI, then there must be cases where m-internal differences *do* lead to differences in evidence. mEI* and dEI* provide clear versions of mEI and dEI that *do* have this consequence. So, while I recognize that mEI* and dEI* may be flawed, I will still use them in the cases that I will describe below. I do not think that any of my examples exploit the problem just mentioned with the formulation.[3]

Given this preparatory work, we can give an argument against mEI* that is structurally identical to the argument Silins gives against EE. Here's how that goes:

Scenario 1

Good Case:

AG = "it appears to me that there is a gorilla walking across the basketball court." At *t* it really does appear to George as if there is a gorilla walking across the basketball court, so *AG* is true of George. However, George doesn't notice

[3] It is worth noting that Silins makes a similar assumption when giving his argument against EE, and for a similar reason.

the gorilla.[4] Let *ATG* = "I appear to be told that it appears to me that there is a gorilla." At *t* it does appear to George that he is told by a neuroscientist that it *did* appear to him as if there was a gorilla. So *ATG* is true of George. George believes that he appears to be told this, but isn't fully confident that it did appear to him as if there was a gorilla. His beliefs are like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

Bad Case:

At *t* it doesn't appear to Bertrand as if there is a gorilla walking across the court. It does, however, appear to Bertrand that he is told by a neuroscientist that it *did* appear to Bertrand as if there was a gorilla. Bertrand believes that he appears to

[4] One might wonder how this is possible. Simons & Chabris ([1999]) demonstrated this phenomenon, which they call 'inattentional blindness'. In the study, the subjects are asked to count the number of passes that the basketball players make during a short video. While the players are passing the basketball, a person in a large gorilla suit walks through the court, beats his chest, and then walks away. Surprisingly, subjects do not notice the gorilla. However, the subjects were not *literally* blind in any kind of plausible sense. It is plausible that it still appears to them as if there was a gorilla, but this appearance simply goes unnoticed. This is bolstered by the fact that in cases very similar to cases of inattentional blindness, *some* information about the unnoticed object does make its way to the agent. Consider this summary of a study by Simons, et. al.:

> …researchers have demonstrated that people fail to notice changes in the information that is visually available to them (Simons 2000). Interestingly, people often cannot describe the change that has taken place, but do demonstrate traces of memory of what they saw before the change. For example, an experimenter holding a basketball stopped pedestrians to ask for directions (Simons et al. 2002). While the pedestrian was giving directions, a group of people (confederates in the experiment) walked between the experimenter and the pedestrian. During this interruption, the experimenter handed the basketball to one person in the group. After giving directions, the pedestrian was asked if he or she noticed any sort of change during the brief exchange with the experimenter. Most did not. However, when led to think about a basketball, the pedestrian did recall seeing it at the beginning of the exchange, and some even recalled specific features of the ball. So, while the participants failed to explicitly notice that a change took place, they did hold accurate implicit memory representations of both the pre- and post-change image. (Chugh & Bazerman, [2007], p. 5)

I take it that this is some evidence that it can *appear* to one as if something is the case (e.g., a basketball was handed off), even if one doesn't notice this. The fact that there are accurate memories of the change imply that it did appear as if something changed, even if the agent didn't notice this.

be told this but isn't fully confident that things appeared this way. He is a d-internal duplicate of Gary, so his beliefs are like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

I think it is quite natural to think that Bertrand and George are epistemically on par in this situation. For both maintain what seems to be sensible doubt about whether or not they were appeared to in a certain way. But even if one doesn't grant this, at least the following is true: someone who thinks that Gina isn't doing worse than Beth shouldn't think that George is doing worse than Bertrand here. For, intuitively, George doesn't have *AG* as evidence. That evidence is not available to him and he maintains reasonable doubts about this appearance.

The problem for the defender of mEI* is that mEI* says that George *is* doing epistemically worse than Bertrand. Why is this? Because mEI* will say that both *AG* and *ATG* are part of George's evidence. After all, it does appear to George that there was a gorilla and it does appear to him as if he's told that it appeared to him that there was a gorilla. Since *AG* is part of his evidence it should be that $cr_t(AG) = 1$. Nevertheless, $cr_t(AG) = 0.9$. Bertrand, on the other hand, is doing perfectly well according to mEI*. For the same reason as it is for George, *ATG* is evidence for Bertrand, and Bertrand conditionalizes on this appropriately. So, mEI* says that George is doing epistemically worse than Bertrand.

Against this, the adherent of dEI* can claim that George and Bertrand are epistemically on par since they are doxastic twins. Thus, the argument goes, dEI* gives the sensible result. George and Bertrand are doxastic duplicates. They have the same

doxastic state, so they have the same evidence. The fact that it *actually* appeared to George as if there was a gorilla is irrelevant. The adherent of mEI* could, of course, reject that in this case *AG* and *ATG* are evidence for George, but as mentioned above, if mEI* is an interesting thesis, there are bound to be cases where things like *AG* and *ATG* are evidence.[5]

One might object that since in Scenario 1 George didn't *notice* that it appeared to him that there was a gorilla walking across the court, it didn't really appear to him that there was a gorilla. This response claims that it appearing to George that there was a gorilla isn't something George shares with his m-internal twins unless George notices that there was a gorilla. So, the objection goes, it appearing to George that there was a gorilla but without his noticing it, *doesn't* make *AG* a member of George's evidence set. If there's really to be a successful argument like this against mEI*, the thought goes, we'll need to have a situation where George *notices* that it appears to him that there is a gorilla. We can, however, easily run such a scenario with the same result. Here's how that goes:

Scenario 2

Good Case:

It really does appear to George as if there is a gorilla walking across the court. He even notices the gorilla. However, his noticing is not consciously accessible.

[5] One might worry that I'm assuming that mEI* is what Silins calls a *mentalist conception of evidence*. "According to mentalist conceptions of evidence, one's evidence consists only of one's mental states or facts about one's mental states." (p. 394) It is true that scenario 1 considers only evidence propositions describing mental states or facts about those mental states (e.g., how one is appeared to). But this is inessential to my argument. For instance, I could have made the relevant propositions *G* (that there was a

It then appears to George as if he is told by a neuroscientist that it *did* appear to George as if there was a gorilla, and that he *did* notice it. George believes that he appears to be told this, but isn't fully confident that it did appear to him as if there was a gorilla. His beliefs are like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

The Bad Case is then similar to the one above, but with the appropriate changes. I take it that this case has just as much force against mEI* as the initial case.

Now, one might object in a very similar way that George noticing that there was a gorilla is not something that George shares with his m-internal twins unless his noticing is consciously accessible. So, the thought goes, if we're really going to make trouble for mEI* with an argument like this, we'll need a situation where George's noticing is consciously *accessible* to him.

Note several things about this kind of response. First, this is clearly *not* the kind of position that many internalists will favor.[6] For internalists have often wanted to avail themselves of *all* parts of the agent's mental state, and not just the accessible parts of the mental state. Any such internalist, then, will need a response to this argument. But, and here's the *tu quoque*, any such defense looks like it will be amenable to the defender of EE against EI. For to respond to these arguments, one must say that there are situations where a subject has evidence *P* because some relation between the subject

---

gorilla on the court) and *TG* (that I was told that there was a gorilla), where *G* is evidence because *AG* is true of George and *TG* is evidence because *ATG* is true.

[6] This includes Silins, who is clear that he does not intend his version of Evidential Internalism to depend on conscious accessibility or access: "Evidential Internalism is not what we may call an *access thesis* in epistemology […] Others might use different readings of "internal", and propose narrower supervenience bases for two thinkers to have the same evidence. For example, an internalist might say that "internal" mental states are just those non-factive mental states which are consciously accessible…" (p. 377)

and *P actually* obtains, even if doubt about this seems appropriate. The mere fact that for mEI* these relations are internal to the agent, and for EE they need not be, is irrelevant.

Second, note how this argument pushes one to more and more internal formulations of evidential internalism. We started out with the thesis that agents with identical mental states don't differ in evidence. Then we were pushed to claim that agents with identical mental states *could* differ in evidence so long as they don't differ in what mental states they noticed. Now, we are considering the claim that agents that notice the same mental states could differ in evidence so long as they don't differ in what is consciously accessible to them.

Finally, note that even this iteration of evidential internalism, in terms of conscious *accessibility*, isn't going to work. For, we can run the argument again even if George's noticing is consciously *accessible*. All that seems crucial to the example is that in the good case George is not actually *consciously accessing* of his noticing. Access*ibility* alone has nothing to do with it. Consider:

Scenario 3

Good Case:

It really does appear to George as if there is a gorilla walking across the court. This is consciously accessible, but George is not conscious of it. It then appears to George as if he is told by a neuroscientist that it *did* appear to George as if there was a gorilla. George believes that he appears to be told this, but isn't fully confident that it did appear to him as if there was a gorilla. His beliefs are like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

Bad Case:

It doesn't appear to Bertrand as if there is a gorilla, so it is not consciously accessible to him. However, it appears to Bertrand as if he is told by a neuroscientist that it *did* appear to Bertrand as if there was a gorilla. Bertrand believes that he appears to be told this, but isn't fully confident that it did appear to him as if there was a gorilla. His beliefs are like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

Since mEI* says that George has more evidence than Bertrand, it implies that George is doing epistemically worse than Bertrand. But this seems implausible. George *isn't* doing worse in this situation. dEI*, on the other hand, doesn't have this consequence. I think this shows that most versions of EI are in exactly the same boat as EE. So if you are attracted to a view like Silins's mEI, then in this respect, you face the same problems as the defender of EE.

4.2.1.4 Accounts Immune to the Tu Quoque

However, this discussion suggests a *different* version of EI that may be immune from this kind of argument. Perhaps, the thought goes, the adherent of EI should say that A and B are internal twins just in case they have the same *conscious* mental states. Then, defend:

cEI*: Necessarily, A and B are conscious internal twins *iff* they have the same evidence.

It is unclear if an argument like Silins's can be given against cEI* by an adherent of dEI*. The relevant kind of case would have to be one where George and Bertrand have all the same doxastic states, and yet George is conscious of something that Bertrand is not. Here is the scenario:

Scenario 4

Good Case

It consciously appears to George as if there is a gorilla. In virtue of this, *AG* is evidence for George. It also consciously appears to George that he is told that it appears to him that there is a gorilla (*ATG*), and he fully believes this. George isn't completely confident, however, that it really appears to him that there is a gorilla. So, his belief state is like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

Bad Case

It does not consciously appear to Bertrand as if there is a gorilla. However, it does consciously appear to Bertrand that he is told that it appears to him that there is a gorilla (*ATG*), and he fully believes this. Although Bertrand fully believes he is told this, he isn't completely confident that given this, it really did appear to him in that way. So, his belief state is like this:

$cr_t(AG) = 0.9$ $cr_t(ATG) = 1$ $cr(AG|ATG) = 0.9$

Now, we must ask: is George doing epistemically worse than Bertrand? If the answer is *no*, then dEI* wins. cEI* entails that George *is* doing worse, so if he isn't, then cEI* is mistaken. But, one might think that in this situation, we finally have a situation where it

is plausible to claim that George is doing worse than Bertrand. After all, it does consciously appear to George as if there is a gorilla. Perhaps he should fully believe this. If that's the right verdict in this case, then perhaps there is a form of evidential internalism, built upon *conscious* experience, that does not succumb to Silins's argument. That is, according to this form of evidential internalism, the agent with less evidence can be doing epistemically better, and yet there is no inclination to say that such a verdict is mistaken.

It is important to note, *why* a view like this might evade the argument. If you think the argument against cEI* fails, it is because you think that if it consciously appears to George as if *P*, then George is fully justified in believing that it does appear to him as if *P*. The guiding thought here is that doubt about this is not epistemically appropriate. Accordingly, we agree that George *is* doing something incorrect by not giving *AG* full credence. So he's doing worse. But if one has more fallibilist intuitions, then even cEI* will not avoid the argument. For one might think that even conscious appearances can be doubted. If so, then we can construct a case like the following:

> It really does appear to George as if there is a gorilla walking across the court. He is conscious of this. But he is sometimes mistaken about how things appear to him, especially when the things in question are odd events – like gorillas walking across basketball courts. In particular, sometimes he is (non-factively) conscious that it appears to him in a certain way, even though it doesn't appear to him in that way.

If cEI* is right, then *AG* is evidence and George should treat this as evidence. But one might feel uncomfortable with George treating *AG* as evidence in this situation. To say

that George shouldn't treat *AG* as evidence in this situation is to open one up to a Silins-style argument. For George will have a doxastic twin, Bertrand, that *doesn't* treat *AG* as evidence, and to whom it doesn't consciously appear as if there is a gorilla. Bertrand will have less evidence, and so be in the bad case, but doing epistemically better than George. On the other hand, to say that George *should* treat *AG* as evidence is to give an account of evidence structurally similar to externalism.

But let's assume that something like cEI* can defend itself from these type of worries. Then we would have one form of evidential internalism, cEI*, that does not succumb to Silins's argument. Note that there is another form of evidential internalism that does not succumb to Silins's argument: dEI*. According to this view, one's evidence is wholly determined by one's doxastic state, and nothing else. Consider one particular instance of such a view according to which *E* is evidence just in case cr(*E*) = 1.[7] One cannot give a Silins-style argument against this account of evidence. Why not? To give the argument we need a scenario where the bad case is judged epistemically better than the good case. This happens when the agent in the good case has more evidence than in the bad case, but his doxastic state does not respond to it. But if it is one's credence in *E* that determines whether or not *E* is evidence, then it isn't possible for the agent's doxastic state not to respond to his evidence. So with such a view, we can't construct the scenario needed for the argument. Thus, we have two theses—cEI* and dEI*—that may evade the arguments above.

However, these two theses both seem to have problems of their own. Consider dEI* first. This is just the deflationary account of evidence considered and criticized in

[7] Recall that this is the kind of view is suggested by Howson & Urbach ([1993]).

Chapter 2. Consider, then, cEI*. I see two problems with cEI*. First, according to this account of evidence you must have conscious states before you have any evidence. This is a bizarre result. It means that a complicated robot that gathers data and executes actions, has *no* evidence so long as it does not have conscious states. Second, and more troubling, according to cEI*, we have a very narrow supervenience base for our evidence. It is implausible that such a narrow supervenience base can really do all the justificatory work that we think our evidence *does*. Just think of all the information we have stored in memory, that seems to do justifacatory work, and yet only a small portion of which is conscious at any time. In sum, cEI* presents us with a very restrictive conception of evidence. It isn't clear that this kind of evidence is always—or even *usually*—the kind of evidence that we care about.

There is, after all, a more objective sense of 'evidence' where what evidence an agent has does not depend on subtle details about what happens to be conscious to the agent with whom we are concerned. It is very plausible that it is this more objective sense of 'evidence' that is at issue when we talk about probability based on evidence, or offer formal Bayesian models of belief change. For example, the relevant evidence in a legal trial does not depend solely on how things *consciously seem* to members of the jury. The evidence in a scientific setting surely doesn't *solely* depend on how things consciously appear to the scientists. Certainly this doesn't seem to be true if we think that regular scientific theories and court verdicts are well-supported by the evidence. Conscious appearance may play a role in the story, and it may be an important one, but it need not be the *whole* story.[8] In fact, Silins's own version of Evidential Internalism

[8] Though see Feldman ([1988]) who appears to take the opposite view.

(mEI) is one account of evidence according to which evidence is more objective, and not solely dependent on conscious experience.

Once we notice that there is this sense of 'evidence' not so intimately tied to what is consciously entertained, then an argument like Silins's against EE is unconvincing. It is unconvincing because the defender of *any* account of evidence—including mEI—will have to answer an exactly similar argument to stop the slide to the kind of extreme-internalism about evidence represented by cEI* and dEI*.

4.2.2 The Undercutting Problem

There are three further problems that can be raised for accounts of evidence that are versions of Evidential Externalism.[9] These three problems are similar in many ways to Silins's objection. The first problem is the Undercutting Problem, the second is the Skeptical Problem, and the third is the Bootstrapping Problem. I address the Undercutting Problem first.

The Undercutting Problem can be best illustrated with an example. Suppose I believe that my sister is sitting across the room from me, because I look at her. If someone informs me that there is an indistinguishable clone of my sister in the area, this information seems to undercut my justification for believing that my sister is sitting across the room. To turn this into an objection to EE (and RAE in particular), assume that my sister is sitting across the room from me and that there is in fact no clone. Let *S* = "my sister is sitting across the room". Let us assume that RAE says that *S* is part of my evidence. This is because I look at my sister and the normal visual process is

[9] Roger White raised some of these worries (particularly the Skeptical Problem and the Undercutting Problem) in his presentation at the Brown Epistemology Conference, February 2009.

sufficiently reliable. Since *S* is part of my evidence it follows from conditionalization that:

$$cr_{t1}(S) = cr(S|S) = 1$$

But now imagine that at *t2* I am informed that there is an indistinguishable clone of my sister in the area. This is false, but I receive this suggestion from a reliable source. Intuitively, this undercuts my justification for *S*. Accordingly, it seems as though *S* is no longer evidence, and one would think that my confidence in *S* should go down.

Now, there are interesting and difficult questions about how to understand the phenomenon of undercut evidence. In the next chapter I address this issue in more detail. But to see how this might be an objection to RAE, we can simply note how EI accounts of evidence seem to do better. In particular, consider a version of EI that is *also* committed to a thesis that we can call *Mentalism*. According to Mentalism, only propositions about an agent's mental states are evidence for that agent, for instance, propositions about appearances or seemings. Note that EI does not entail Mentalism. EI is only committed to the claim that internal twins have the same evidence. But if one is committed to EI *and* the thesis that evidence must be true (or very often true), then EI will often give one evidence propositions that are only about appearances or seemings, so EI will is naturally aligned with Mentalism. Further, the attraction of EI often lies in this alignment with Mentalism. So, assume such an account of evidence.

Let *AS* = "it appears that my sister is across the room", and let *TC* = "I appear to be told that there is a sister-clone". Here's how things go according to such an account. Initially:

$$cr(S|AS) = 0.99$$

Then, by looking across the room at *t1*, I don't get *S* as evidence. Instead, I get *AS* and conditionalize. Thus:

$$cr_{t1}(AS) = 1 \qquad cr_{t1}(S) = 0.99$$

Nevertheless, it can still be the case that cr(S|AS ∧ TC) = 0.5. So, when I'm told about the clone at *t2*, and I get *TC* as evidence, my credences are:

$$cr_{t2}(AS) = 1 \qquad cr_{t2}(TC) = 1 \qquad cr_{t2}(S) = cr(S|AS \wedge TC) = 0.5$$

On this picture we see how *S* can be undercut by additional evidence, like *TC*. We lose this capability, so goes the objection, when we treat *S* itself as part of my evidence. Instead, we should treat *AS*—something about how things *appear* to me—as my evidence. If, however, evidence only concerns how things *appear* to me, then it would seem that the members of my evidence set are internally specified.

4.2.2.1 Response

First note that this is not a problem unique to versions of EE, for we can construct a case structurally similar that causes problem for EI. Grant that it does appear to me that my sister is sitting across the room. So, *AS* is part of my evidence, though not *S*. Since *AS* is part of my evidence it follows that:

$$cr_{t1}(AS) = cr(AS|AS) = 1$$

But now imagine that at *t2* I am informed by a reliable source that there is a faulty connection between my credences and the way things appear to me. "Sure," the source tells me, "you *believe* that it appears to you as if your sister is sitting across from you, but that doesn't mean it actually appears to you that way." The reliable source need not be telling the truth. Perhaps it *does* appear to me that my sister is sitting across the room. All it takes is doubt that this is so. Intuition says that this undercuts my

justification for *AS*. Accordingly, it seems as though *AS* is no longer evidence, and one would think that my confidence in *AS* should go down.

According to EE, one's evidence is determined by the appropriate relation obtaining between the state of the world and the agent's doxastic state:

State of the World → Doxastic State

The undercutter in the first story questioned whether or not this relation holds. According to MSE, on the other hand, the members of my evidence set are determined by some relation between my mental state and my doxastic state:

Mental State → Doxastic State

The undercutter works in exactly the same way, questioning whether or not the relation holds. So, we get a problem analogous to the problem for EE. But this is a problem for an *internal* account of evidence. Further, in this case we can't appeal to appearances to explain why this undercutting defeater has the effect it does. For we're granting that it does appear to me that she is sitting there. It is just that I get reason to believe that how things appear to me aren't hooked up with my doxastic state correctly.

There are several responses the defender of EI might make. First, the defender might say that *AS* really is part of my evidence if it actually does appear to me that *S*, even if I receive undercutting evidence. But this is just to deny that there could be any undercutting in the way just described, without explaining *why* this is so. One possible explanation is that *AS* is part of my evidence just in case there *is* an appropriate connection between the appearance and cr(*AS*). The problem with this, however, is that it looks like a variant of externalism about evidence. The externalist says that *S* is evidence just in case there *is* an appropriate connection between my sister being across

the room and cr(*S*). Both share the feature that some non-believed relation obtaining renders a proposition evidence in an non-undercuttable way. The fact that one of these relations concerns states inside the head of the agent does not seem to be of any epistemological relevance.

A different response is to define 'appearance' in such a way that appearances are the kinds of things about which doubt is rationally impossible. According to this response, it is impossible that the undercutting described above ever occurs rationally, for if it does, that only shows that it didn't really appear to me that my sister was sitting across the room. The problem with this response is that it sets the bar too high for what it is to be appeared to in some way. For as long as beliefs about appearances are distinct from appearances, it always seems possible to rationally entertain doubt about the connection between the two.

In summary, then, EI and EE both face versions of the Undercutting Problem. Given this, the possibility of undercut evidence cannot be used in an argument against EE. This kind of response to the problem of undercutting shows that the Undercutting Problem is not a problem for EE alone. However, a more positive response would be nice. Though I haven't given such a response here, in Chapter 5 I will show that RAE has the resources to model this kind of undercutting in a satisfying way, whether or not it is an internal or an external proposition that is being undercut. The basic idea is that if it is the reliability of a process indicating a proposition that makes a proposition evidence (that is, if RAE is correct), then strong doubts about the reliability of that process are often sufficient to expunge the proposition from the evidence set. I postpone

until Chapter 5 a detailed discussion of this proposal and how it can be consistent with a reliabilist understanding of evidence.

4.2.3 The Skeptical Problem

The Skeptical Problem brings up issues similar to the Undercutting Problem. Let's say that I've just finished watching *The Matrix* for the first time. I think about the film all night, and as I close my eyes to go to sleep I am four times more confident that I'm actually in the Matrix (M) than not ($\neg M$). So:

$$cr(M) = 0.8 \qquad cr(\neg M) = 0.2$$

Further, because I understand how the Matrix works (I've just seen the film, after all), I'm very confident that I'm not in my room ($\neg R$) given that I'm in the Matrix. Thus:

$$cr(R|M) = 0.05$$

Given all this, it follows that as I fall asleep, my credence that I'm in my room is between 0.4 and 0.24.[10]

Let's suppose I have a dreamless sleep, and none of my doxastic attitudes change. I wake up in the morning (t) and first thing see that I'm in my room. Let's say that my visual processes are reliable enough to make R part of my evidence. So, what happens? Well, $cr_t(R) = 1$. If we assume that $cr(R|\neg M) = 0.95$, then this results in me having credence of only about 0.17 that I'm in the Matrix.[11]

---

[10] Since $cr(R|M) = cr(R \wedge M)/cr(M) = 0.05$, and since $cr(M) = 0.8$, it follows that $cr(R \wedge M) = 0.04$. Since $cr(\neg M) = 0.2$, it follows that $cr(R \wedge \neg M) \leq 0.2$. Since $cr(R) = cr(R \wedge M) + cr(R \wedge \neg M)$, it follows that $0.04 \leq cr(R) \leq 0.24$.

[11] $cr_t(M) = cr(M|R) = cr(M \wedge R)/cr(R) = cr(M \wedge R)/[cr(M \wedge R) + cr(R \wedge \neg M)]$.From above, $cr(M \wedge R) = 0.04$. Given our assumption, $cr(R|\neg M) = 0.95$, and so $cr(R \wedge \neg M) = 0.95cr(\neg M)$. Since $cr(\neg M) = 0.2$, $cr(R \wedge \neg M) = 0.19$. Plugging this all in, $cr_t(M) \approx 0.17$.

This strikes many as counterintuitive. Before going to bed I was much more confident than not that I was in the Matrix, and yet reliably looking at my room dramatically drops my confidence that I'm in the Matrix to less than 1/5. This doesn't seem to be the right way to understand what should happen in skeptical scenarios.

Again, a different Mentalist account of evidence seems to do better. If my evidence in the morning isn't *R*, but rather that it *appears* as if *R* (*AR*), then we seem to be able to better model intuitions here. Presumably, if I'm very confident that I'm in the Matrix, then it appearing as if I'm in my room isn't going to shake my confidence that I'm in the Matrix at all, for that's exactly what I'd expect if I were in the Matrix. This is exactly why the Matrix scenario (and other similar skeptical scenarios) are so maddening. So, even though upon waking up, $cr_t(AR) = 1$, it is plausible to maintain that $cr_t(M)$ is still high, since $cr(M|AR)$ is high. Thus, upon waking up, we aren't forced to conclude that I must drastically lower my confidence in *M*.

4.2.3.1 Response

Again, it is important to note that there is an analogue of the skeptical problem for an adherent of EI. Let's say that I've just finished watching a documentary about psychological studies showing how we often falsely believe things about how things appear to us. In particular, let's say that the documentary casts doubt on one's accuracy about numerical appearances, where a belief about a numerical appearance is of the form: "it appears to me as though there are *n* Xs" (*AnX*). I think about the documentary and as I lay down to go to sleep, I'm four times more confident that my beliefs about numerical appearances are inaccurate. Call this hypothesis, *F*. So:

$cr(F) = 0.8 \qquad cr(\neg F) = 0.2$

Further, because I understand what *F* is saying (I just saw the documentary, after all), I am such that for any proposition of the form *AnX*, cr(*AnX*|*F*) = 0.05. Accordingly, for any proposition of the form *AnX*, $0.4 \leq \text{cr}(AnX) \leq 0.24$.

Now, I have a dreamless night, with none of my doxastic states changing. I wake up in the morning (*t*) and first thing it appears to me as if there are four spots on the ceiling (*A4S*). The internalist who ties evidence to appearances says that *A4S* is part of my evidence. So, $\text{cr}_t(A4S) = 1$, which, granting the same assumptions as above, has the effect of instantly making my credence in *F* approximately 0.17.

I think that this is just as counterintuitive as the Matrix case above. I'm very confident that certain beliefs about how things appear to me are inaccurate. However, once I have a numerical appearance, I am rationally required to be certain that I have such an appearance, and as a result reduce drastically my credence that my beliefs about how things appear to me are inaccurate.

Now, I'm not saying that were I to change my beliefs in this way, I would not in some sense making the *right* response. It *does* appear to me as if there are four spots on the ceiling, and I *do* believe it. There is something right about this. My point is that this is right in just the same sense that it is right for me to believe that I *am* in my bedroom after seeing the Matrix. I have as evidence that I'm in my bedroom, and I *do* believe it.

There is a slight complication with the scenario I've just given. In particular, one might worry that it doesn't make sense to say, in general, that cr(*AnX*|*F*) =0.05. This is because *AnX*, as an appearance proposition, is an indexical proposition essentially referring to *now*. It says something like:

"It appears (to me) as if there are *n* Xs **now**."

But if cr(*AnX*|*F*) = 0.05, then if I were to learn *F*, I would be rationally required to be such that for all the *AnX* propositions, cr(*AnX*) = 0.05. But this seems absurd. Presumably, for most *AnX* propositions cr(*AnX*) should be 0 (or *very* close to 0). What we wanted to maintain is that $cr_t(AnX)$ should be 0.05 for those numerical propositions that describe how it appears to me at *t*.

This is indeed a problem with how I've presented the case.[12] But it does not show that the skeptical problem is not a problem for the internalist about evidence. What it shows is that there is another puzzle about how to best represent this kind of skeptical scenario if one is a Bayesian. But it is obvious that there is just as much a puzzle about such skeptical scenarios whether one is an internalist or an externalist about evidence. Internal matters no less than external matters can be rationally doubted. Internalism about evidence, then, is in the same boat as externalism with respect to this kind of skeptical problem.

This response to the Skeptical Problem is enough if one wants to defend versions of EE against EI. Again, however, such a response simply shows that the Skeptical Problem is not the externalist's alone. It doesn't show that RAE can handle such a case. To see how this might be done, note first that the Skeptical Problem is very similar to the Undercutting Problem, except that here we have a case of "preemptive" undercutting. The agent has high credence in an undercutter for a certain proposition

[12] This general phenomenon is discussed in Weisberg [*forthcoming*]. Weisberg diagnoses the problem as one stemming from Rigidity. I devote the next chapter to considering the issue in detail. Note that we could get around the problem by positing a more basic set of propositions. Call these *seemings*. Let *SAnX* refer to the proposition that it seems that it appears as if there are *n* Xs. The idea is that the effect that *F* has on me is not to make it so that cr(*AnX*|*F*) = 0.05, but rather that cr(*AnX*|*F* ∧ *SAnX*) = 0.05. Then, when I wake up it *seems* that it appears to me as if there are four spots on the ceiling. Given my confidence in the skeptical hypothesis (*F*), it seems as though my credence that it actually appears to me in this way should be low. However, if EI follows something like MSE then *AnX* will be evidence, and so we will get

(describing the skeptical scenario) *before* considering that proposition. Given this structural similarity, the account of undercutting to be sketched in Chapter 5 is relevant to the Skeptical Problem.

But even without such an account note that RAE seems to have the resources to model this kind of situation. RAE, recall, is formulated with a variable reliability threshold for when a proposition becomes evidence. We will, then, often be able to model skeptical scenarios in the way desired by simply raising the level of reliability required for a proposition to be evidence. Since, for most of us, most of the time, we have a more reliable route to propositions about how things appear to us than about how things *are*, raising the level of reliability will permit propositions like "It appears that I'm in my room" as evidence while disallowing propositions like "I am in my room." This will allow us to model the skeptical situations as desired.

As a final response to the Skeptical Problem, we can note that modeling extreme skeptical scenarios in a completely satisfactory manner may actually involve a different notion of evidence, more along the lines of guidance, rather than along the lines of an objective, evaluative notion that I sketched in Chapter 3. The defender of RAE, then, can happily claim that extreme skeptical scenarios are not those that RAE is meant to capture. Nevertheless, RAE can be made to work in many such situations by altering the reliability threshold.[13]

---

the unintuitive result just as we did for EE. This follows closely the dialectic that emerged when discussing Silins's argument.

[13] Note: RCE would allow one to avoid this entire discussion. For RCE, as a necessary condition on having evidence, could be paired with some condition which guarantees that agents can't get anti-skeptical evidence (e.g., "I have hands") in skeptical scenarios. Thus, even a reliably indicated anti-skeptical proposition couldn't be evidence in such a situation. Now, a condition like this strikes me as too stipulative. It is preferable, I think, to use RAE and get as much mileage out of reliability as possible towards modeling such situations. But it is worth noting that backing off to the weaker RCE would allow one to avoid the Skeptical Problem.

4.2.4 The Bootstrapping Problem

A final objection claims that externalism about evidence will be subject to a version of the Bootstrapping Problem. Bootstrapping happens when an agent is able to manufacture justification for some proposition, seemingly out of thin air. Consider the classic example of bootstrapping[14]: in looking at my gas gauge I conclude that my tank is full. I also conclude that the gauge reads "Full". If I do this many times, it looks as if I can conclude, just from looking at the gas gauge, that the gas gauge is perfectly reliable. This strikes one as inappropriate epistemic behavior.

Here's how the problem goes for the proponent of externalism about evidence[15]: There is a lion behind the tree, and one's visual system is quite reliable with respect to lions. Accordingly, the defender of RAE says that *L*: "there is a lion behind the tree" is evidence. Suppose, however, that one is also quite reliable in one's beliefs about how things appear, so that *AL*: "there appears to be a lion behind the tree" is evidence, too. But then, given conditionalization:

$$cr_t(AL) = cr_t(L) = 1$$

From this it follows that $cr_t(AL \wedge L) = 1$. If the agent then has many similar encounters, it seems that he can then easily conclude that how things appear are perfectly correlated with the way that they are. We seem to be saddled with the following problem: if the relevant perceptual faculties *are* reliable, then agents too-quickly and too-easily become confident that they are reliable.

Here's how an alternative picture seems to avoid this. If only *AL* is evidence, then only $cr_t(AL) = 1$. If , say, $cr(L|AL) = 0.99$, then $cr_t(L) = 0.99$. From this it follows

[14] Vogel ([2000]).

that $cr_t(AL \wedge L) = 0.99$. This doesn't look much better than the alternative picture given above. However, there is a significant difference. Since the conjunction of *AL* and *L* is given less than maximal credence, when it is conjoined with many other such conjunctions, the credence for the whole conjunction can fall dramatically. Accordingly, it doesn't seem as if the agent can conclude that the way things appear are the way things are. Further, if one objects to the very first step, where just as a result of looking across the room $cr_t(AL \wedge L) = 0.99$, one can claim that $cr(L|AL)$ was set at too high a value. There is no such option if both *AL* and *L* are part of the evidence set, since independent of the value of $cr(L|AL)$, $cr_t(AL) = cr_t(L) = 1$.

4.2.4.1 Response

As with the other objections, there is a clear *tu quoque* response that could be made, where both propositions are evidence in virtue of *internal* features of the agent. Since it is clear how such a response would go, I will not present it explicitly. This neutralizes the threat to the Evidential Externalist from the Internalist. However, we would still like some sort of positive response. I will pursue two responses. First, following an argument made by Hilary Kornblith ([2009]), I will argue that the adherent of RAE need not be worried that any proposition like the following ends up being *evidence*:

C: "My lion appearances are perfectly correlated with the presence of lions."

This neutralizes the force of the Bootstrapping Objection. However, even given this response, RAE will sometimes say that conjunctions like $(L \wedge AL)$ should be given full credence. Of course, this is unobjectionable if there were different routes to *L* and *AL* respectively. However, it looks particularly objectionable if *L* was *inferred* from *AL*.

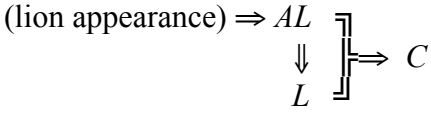
---

[15] I take this example from Weatherson ([*ms*]).

Thus, my second response will be to consider these objectionable cases, and then offer and motivate a substantive addition to RAE. First, however, I will deal with the main Bootstrapping Problem.

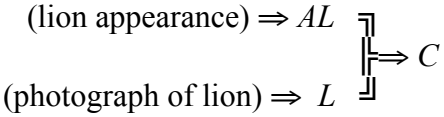
In his ([2009]), Kornblith discusses the phenomenon of bootstrapping and notes that the bootstrapping process of coming to believe a proposition like *C* can be very unreliable. Since this is what matters for the reliabilist, the reliabilist need not say that bootstrapping is acceptable. Kornblith's response asks us to look at the entire process that leads to the belief that *C*. This is a process that could be schematically represented like this:

Generic Bootstrapping (GB)



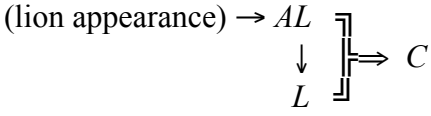
where the arrows indicate the cause of the beliefs in question. To get a feel for the diagram, this could be contrasted with a non-bootstrapping way of coming to believe *C*:

Regular Checking



Now consider a special kind of bootstrapping process. Let an single line arrow denote not just a process of belief production but rather a *reliable* process:

Specific Bootstrapping (SB)



If we want to know if the bootstrapping process of coming to believe *C* yields a justified belief according to reliabilism, we need to know what type of process results in the belief that *C*. So, if we want to know if the belief that *C* is justified, we need to know something about the rest of the bootstrapping inferences that the agent in question draws, and not just the reliability of the "components" that make up the entire process. But there is an ambiguity. Generic bootstrapping is clearly not going to be a reliable process of belief formation. Specific bootstrapping, on the other hand, will be. So, if we want to know if some agent is justified in believing *C* we need to know what kind of process leads to that belief. The intuition against bootstrapping comes from thinking of the process of belief formation as being of the GB-type. But reliabilism can agree that if *C* is formed via a GB-type process of formation, then *C* is formed in an unreliable way. This is so even if in *this particular instance*, the components of the GB-type process are reliable.

How does this apply to RAE? It may be that there is a reliable process indicating *L* and *AL*. But this does not mean that the agent has a reliable process indicating *C*, especially if the agent in question makes bootstrapping inferences all the time. So, the defender of RAE need not say that a proposition like *C* is *evidence*.

This takes the sting out of the Bootstrapping Objection. However, it is only a partial response. For RAE *will* say (in certain circumstances) that *L* and *AL* are evidence, and so condone giving full credence to $(L \wedge AL)$. Some might find it

objectionable that even one proposition like ($L \wedge AL$) can be given full credence in the way described. In the next section, I will consider this issue in more detail.[16]

4.2.5 Summary

I have considered four objections to RAE that are based on the fact that RAE is a species of Evidential Externalism. I have noted that all four objections arise in analogous ways for Evidential Internalism, and so pose no direct threat to RAE. However, I have also offered more specific responses concerning how RAE can handle and correctly model the situations introduced by the objections.

4.3 Objections to RAE

In the last section we considered arguments that took their aim at RAE because it is a species of Evidential Externalism. In the next section, I consider arguments that take aim at RAE in particular. Some considerations will motivate additions to RAE, while I will argue that RAE already has the resources to respond to some of the others.

4.3.1 Evidence and Inference

Consider a case where the agent gets both L and AL as evidence, and so ($L \wedge AL$) receives full credence. Of course, it is not *always* inappropriate to have full credence in both L and AL. The most obvious cases of this occur when each proposition is indicated

[16] One might think that there is another problem: if the agent is such that for large n, $\text{cr}(C|(L_1 \wedge AL_1) \wedge (L_2 \wedge AL_2) \wedge \ldots \wedge (L_n \wedge AL_n)) = high$, then it follows from RAE that the agent can justifiably become very confident in C, purely from looking at lions Note, however, that in most cases if someone embarked on such a project of confirming C, this wouldn't work. For, as one looks at lions, one would also get evidence that one has embarked on a project of confirming C via bootstrapping. It is plausible, however, that the value of $\text{cr}(C|(L_1 \wedge AL_1) \wedge (L_2 \wedge AL_2) \wedge \ldots \wedge (L_n \wedge AL_n) \wedge$ **I'm bootstrapping**) is low. Further, RAE is particularly well-equipped to argue that some proposition like *I'm bootstrapping* is evidence in these cases.

by a distinct process. For example, perhaps *AL* is rendered evidence via some process and *L* is rendered evidence via some wholly distinct process. This is analogous to the situation where you check the gas gauge and then independently check the amount of gas in the car. There is nothing inappropriate about this.

The cases that seem worrying are ones where one of *L* or *AL* is inferred from the other. So, perhaps the agent comes to believe *AL*, and then from this infers *L*. Or, perhaps the agent comes to believe *L* and then from this infers *AL*.[17] If there is a scenario in which it is inappropriate for *L* and *AL* to both be evidence, it is this situation. Certainly, there have been those that have registered intuitions that this is the case.[18] Further, if this is a problem, then it is a problem for RAE, since RAE says that *any* reliable (enough) process of belief-formation can yield evidence. I do not think that an appeal to intuitions is sufficient to decide whether or not there is anything wrong with getting some proposition as evidence via inference. Nevertheless, I will argue that there *is* something epistemically problematic with a situation in which an agent gets some proposition as evidence via inference. This will motivate a substantive addition to RAE to take account of this fact.

4.3.1.1 Bayesian Inference

I will show that if we understand inference on the Bayesian model, then there is reason to say that something has gone wrong if an agent gets evidence from inference. First we must say something about inference on the Bayesian model.

[17] There is some support from cognitive science that this latter description might indeed be a realistic picture of how such beliefs are formed. Lyons [2009] gives a survey of the literature in this area.
[18] For instance, Maher ([1996]), Weatherson ([*ms*]), Littlejohn ([*ms*]).

According to a natural way of interpreting the Bayesian formalism, credences change in two ways:

(1) a credence in a proposition increases because the proposition is evidence,

(2) a credence in a proposition increases/decreases in virtue of a conditionalization update.

The latter of these changes is most like an inference, since it is a change in a credence, brought about solely in virtue of some of the agent's other credences (namely, the agent's credences in the evidence propositions). One artificial feature of the Bayesian framework is that both changes are represented as taking place instantaneously. So, in one epistemic moment, the evidence is received (1), and then this evidence is propagated through the rest of the belief state (2). Nevertheless, epistemologists using the formalism can make a distinction between changes that are solely the result of changes in other beliefs and those that are not. That is, suppose that I look out the window at the cloudy sky and then conclude that it will rain today. Before I look out the window, my credences might be something like:

$\mathrm{cr}_0(C) = 0.5$ $\mathrm{cr}_0(R) = 0.5$ $\mathrm{cr}_0(R|C) = 0.9$

Then, upon looking out the window, my credences change to:

$\mathrm{cr}_1(C) = 1$ $\mathrm{cr}_1(R) = 0.9$

This is naturally understood as describing a scenario where *C* is evidence for me, and then, on the basis of this evidence, I conclude *R*. Now, in this account of the scenario, I do not have full credence in *R*. Nevertheless, it seems correct to say that I do believe that it will rain today.[19] So, we can talk about the belief that it will rain today and the

[19] This is compatible with what I take to be the dominant view that belief does not require credence 1.

belief that it is cloudy. This scenario is best described as one where the belief *R* is formed via a process of inference, whereas the belief *C* is not formed via a process of inference. Further, it seems appropriate to characterize this situation as one where *R* is inferred from the belief that it is cloudy. Thus, on the Bayesian picture, an inference is a change in a credence that is the result of a conditionalization update.[20]

The question before us, then, is the following:

> Is it possible for *R* to be evidence, given that the belief that *R* was acquired via an inference from the belief that *C*?

We want to know if it is possible for an agent to get evidence via a conditionalization update. My answer will be that whenever an agent gets evidence in this way, something has gone epistemically wrong.

4.3.1.2 Inference from a Proposition Treated as Evidence

My argument has two parts. In this first part, I will consider the possibility that *R* is evidence in virtue of being inferred from *C*, where *C* is itself treated as evidence by the agent. This fits in well with my characterization of Bayesian inference, where an inference is the result of a conditionalization update on some evidence. After arguing in this section that this is problematic, I will move to the second part of the argument. In the second part, I will consider the possibility that *R* is evidence in virtue of being inferred from *C*, but where the agent does not treat *C* as evidence.

---

[20] This is a tricky issue. For suppose that *C* is evidence. Then, one can see the agent's credence in *C* as a result of conditionalizing on *C* and the fact that $cr(C|C) = 1$. Thus, on this picture, *every* change in the agent's credence is the result of a conditionalization update. One can interpret the formalism in this way. But if one does, then we lose the distinction between changes in the doxastic state brought about by other beliefs and changes brought about by something else. On this picture, every change in a doxastic state is brought about by something other than a belief. Since inference is a change from a belief to another

Suppose, then, that the agent treats C as evidence, so assigns C full credence. Suppose further that as a result of this, the agent infers R, and that R thus becomes evidence. Since R is formed via a Bayesian inference from C, it must be that $cr_0(R|C) = b$, where b is some high credence value, high enough so as to count the resulting credence in R as a belief. Consider first the case where $b < 1$.

Let's think through this situation. In virtue of the fact that C is evidence and $cr_0(R|C) = b$, the agent executes an inference, the result being $cr_1(R) = b$. According to the view being considered, R is evidence in virtue of this inference. Given that R is evidence, however, it follows from *COND* that $cr_1(R)$ should be 1. So, if the agent updates via conditionalization in this case, he is doing something inappropriate: $cr_1(R) = b$ and yet it should be 1.

Suppose, then, that instead of $cr_1(R)$ being assigned a value of b, it had been assigned a value of 1. If it had been that $cr_1(R) = 1$, however, then the process leading to the belief in R wouldn't have been a Bayesian inference from C. But it is just that inference in virtue of which R is supposed to be evidence. So, if R had been treated as evidence in virtue of a Bayesian inference, R wouldn't have been the result of a Bayesian inference. So, if $b < 1$, there is a problem with acquiring evidence via a Bayesian inference.

This part of the argument relies on a particular picture about what an inference is. According to this picture, an agent's doxastic attitude towards some proposition Q is the result of an inference from some set of propositions **P** only if the credence the agent assigns to Q is equal to $cr(Q|\mathbf{P})$. One could challenge this claim by providing some

---

belief, this interpretation of the formalism does not allow us to say anything about inference. So, I think it best to see changes in evidence propositions as not resulting from a conditionalization update.

other model of inference within the Bayesian scheme. However, I think this would be a challenging thing to do. For suppose that we think of an agent's doxastic attitude towards some proposition Q as the result of an inference from some set of propositions **P** just in case that doxastic attitude is *caused* by the agent's doxastic attitude toward the set of propositions **P**. This alternative picture allows that an agent's doxastic attitude towards Q is the result of an inference from some set of propositions **P** without the credence the agent assigns to Q being equal to $\mathrm{cr}(Q|\mathbf{P})$. However, this picture does not distinguish genuine inference from mere causation. For certainly beliefs can be the cause of other beliefs without being *inferences*. Appealing to the value of $\mathrm{cr}(Q|\mathbf{P})$ is one way to make clear that the *content* of the propositions in the set **P** are important to the inference.

So far we have considered the case where $\mathrm{cr}_0(R|C) = b$ with $b < 1$, and the agent updates on evidence C. We saw that in such a case, it is problematic to think that R itself could become evidence. Consider, now, the case where $b = 1$. If $b = 1$ then when the agent conditionalizes on C in accordance with $\mathrm{cr}_0(R|C) = b$, the effect is that $\mathrm{cr}_1(R) = 1$ in which case R is treated as evidence. Above I complained that if the agent gets R as evidence by inferring it from C then R is not treated as evidence. But this complaint does not hold if $b = 1$. So there is only a problem on the assumption that $b \neq 1$. What is the import of this? It is plausible to think that the difference between the cases where $b = 1$ and where $b \neq 1$corresponds to the difference between whether or not the agent sees himself as making a deductive or an inductive inference. My argument shows that there

is something inappropriate with getting evidence via inductive inference, but not via deductive inference.[21]

So, the argument so far has shown that there is something inappropriate with inductive inferential evidence, where the propositions from which the inference is drawn are themselves treated as evidence. Nothing has been shown to be mistaken with deductive inferential evidence in such situations. In the next section, I consider inferences drawn from propositions that are not themselves treated as evidence.

4.3.1.3 Inferences Drawn from a Proposition Not Treated as Evidence

Above I assumed that the agent treats the proposition from which the other belief is inferred as evidence. That is, the agent treats C as evidence and then infers R from this. But suppose that R is inferred from C, but C is not itself treated as evidence by the agent, so that $cr(C) \neq 1$. What should we say about this situation? Note, first, that in this case we simply do not have a conditionalization update, and so we do not obviously have something that should be described as an inference from C to R. But suppose we just say that R is evidence at t_1 in virtue of the fact that it was inferred from $cr_1(C) = n < 1$. I grant that it is hard to know just what this means, but I will show that there is a problem with this nonetheless.

Now, either $cr_0(C|R) \neq cr_0(C)$ or $cr_0(C|R) = cr_0(C)$. Suppose first that $cr_0(C|R) \neq cr_0(C)$. Then, upon getting R as evidence at t_1, the agent must conditionalize on this thus leading to a change from t_0 to t_1 in the credence value assigned to C. That is, the value that $cr_1(C)$ has is at least in part due to the fact that R is evidence at t_1. But then the

[21] Of course, this is not to say that every time an agent sees himself as making a deductive inference from some evidence, that the result counts as evidence. It is just to say that my argument does not show that

value of $cr_1(C)$ depends on R being evidence and yet R's status as evidence depends on it being inferred from $cr_1(C)$. I maintain that this kind of circular dependence is absurd. Certainly, we would not think that it is required epistemic behavior.

Suppose, then, that $cr_0(C|R) = cr_0(C)$. Accordingly, there is no change from t_0 to t_1 in the credence value assigned to C, and so we need not say that the value of $cr_1(C)$ depends on R being evidence. However, if $cr_0(C|R) = cr_0(C)$, then it follows that $cr_0(R|C) = cr_0(R)$.[22] This means that at t_0 the agent does not regard C as evidence for R. But then how, at t_1 does the agent come to infer R directly from C? It doesn't seem as though we should describe the processes that led to the belief in R as an *inference* from C so long as $cr_0(R|C) = cr_0(R)$. So, although we don't get circular dependence when $cr_0(C|R) = cr_0(C)$, we also don't have a situation where R is inferred from C. So, if R is inferred from C, but C is not itself treated as evidence by the agent, R cannot be evidence for the agent without problems arising.

Summing up, I have shown that there is a problem with thinking that evidence can be garnered via a conditionalization update from some proposition that is treated as evidence, and there is a problem with thinking that evidence can be garnered via an inference from some proposition that is not treated as evidence. Perhaps there is a different way of understanding inference according to which these problems do not arise. But if so, I do not know what it is. In addition, it seems clear that any such account of inference would be a significant addition to the traditional Bayesian picture. Thus, I think we have good reason to claim that a proposition inductively inferred as the result of a Bayesian inference cannot itself be evidence.

---

there is something always wrong with evidence garnered in this way.

4.3.1.4 Modification of RAE

Some register strong intuitions that a process of inference cannot yield evidence. The result just obtained fits well with these intuitions. But importantly, these intuitions are not what led to this conclusion. I am pleased about this because I find that in myself the relevant intuitions are far from clear.

Note, however, that if evidence cannot be obtained via inference in this way, then this requires a substantive change to RAE. RAE says that *any* reliable belief-forming process is one that yields evidence. I have now presented reasons for thinking that we should not allow certain kinds of inferential processes to yield evidence, regardless of their reliability. Thus, RAE must be altered. The new RAE, thus says the following:

> **RAE + inference condition:** The set of propositions, $\mathbf{E_t}$ is S's evidence set at *t* iff there are reliable belief-forming processes **(that are not inductive inferences)** available to S at *t* such that if S applied those operations S would believe all the members of $\mathbf{E_t}$ at *t* and those beliefs would be caused by the reliable belief-forming processes.

In what follows I will note how this addition to RAE allows for certain responses to objections that would not otherwise be possible. However, let me now note some things that this addition does *not* mean.

First, the addition does *not* mean that I can't get as evidence facts about my own belief state. That is, there is a difference between inferring some proposition from a believed proposition, and doing something like an internal scan to realize that one believes some proposition. In particular, an internal scan of one of your beliefs is not

---

[22] Suppose $cr(C|R) = cr(C)$. Since $cr(C|R) = cr(C \wedge R)/cr(R)$, it follows that $cr(C \wedge R) = cr(C) \times cr(R)$. Since $cr(R|C) = cr(C \wedge R)/cr(C)$, it follows that $cr(R|C) = cr(C) \times cr(R)/cr(C) = cr(R)$.

the same as doing a Bayesian update: one is an inference, and one is not.[23] So, this addition does not require us to deny that information about our own belief state is evidence.

There is a second thing that the addition does not require. It does not require us to say that having certain beliefs cannot be relevant to the processes that agents have available to them or even the reliability of those processes. For instance, it might be that I am much be much less reliable about a certain domain when in possession of certain background beliefs. For instance, if I strongly believe that the lighting is misleading, I might be far less reliable in some of my visual beliefs. Further, if I strongly believe that the lighting is misleading, I might not even have available certain processes of belief formation that I would otherwise have available to me. The addition to RAE does not speak to these situations. Rather, the addition says that you cannot get evidence solely via an inference from another belief.

4.3.2 The Urn Objection

Consider an example considered by Timothy Williamson ([2000]) that is relevant to RAE. There is an urn filled with balls. I have seen 999 balls removed, and all have been black. Grant that the propositions describing each of these draws is part of my evidence. It seems that I can reliably infer that the 1,000th ball will be black (call this proposition

[23] Suppose, however, that we think of inference as proceeding directly via some proposition. One might wonder why the problem of circular dependence doesn't arise for a case when you scan your beliefs in a similar way that it arose for inference. Suppose that P is not evidence, $cr_0(P) = 0.7$, and that $cr_0(P|<cr(P) = 0.7>) = 0.2$. (I use angled brackets '< >' to denote the proposition *that* a credence takes a certain value.) That is, I believe that I overestimate things like P. Now, it might be that at t_1, I get as evidence $<cr_0(P) = 0.7>$, via some process that depends on $cr_0(P) = 0.7$. Nevertheless, upon getting this evidence, the value of my credence in P may change to $cr_1(P) = 0.2$. Is there any kind of objectionable circular dependence here? I don't think so. It is my *past* belief state at t_0 on which my evidence at t_1 ($<cr_0(P) = 0.7>$) depends, and the value of $cr_1(P)$ at t_1 depends on that evidence.

'$B_{1,000}$'). Thus, RAE would seem to say that it is part of my evidence. Now my evidence includes that 1,000 balls are black. From this it seems that I can reliably infer that the $1{,}001^{st}$ ball will be black. So, $B_{1,001}$ is evidence. In short, it seems that for any n, B_n is evidence for me. But then, it follows from RAE that I am required to be rationally certain of every B_n proposition. Now, after seeing 999 black balls removed by a random process, perhaps I should have a very high degree of belief in $B_{10,000}$. But it doesn't seem as though I should be *absolutely* confident that it is true. Yet RAE appears to have this consequence. This seems to be the sort of case that Williamson has in mind when he writes: "If evidence required only justified true belief, or some other good cognitive status short of knowledge, then a critical mass of evidence could set off a kind of chain reaction." (p. 201)

By way of response one can first note that it is not so bad to make inferences that look very much like the allegedly unacceptable one above. For instance, instead of observing balls pulled from an urn, perhaps I am observing whether when I drop a cannonball from the top of a tower it falls towards or away from the earth. After 999 such observations, we surely think that I am justified in believing that objects freely fall towards (rather than away from) the earth. But even more than this, it is plausible to think that such a proposition is *evidence* for me. In this context, the evidential chain reaction doesn't look so bad. Of course, this is not a complete response, as we are left with the unsettled feeling that something has still gone wrong in the urn case.

In what follows I will give two responses. First, I will explain why RAE does not commit one to the radical conclusion that $B_{10,000}$ is evidence. Second, I will show how adding the inference condition to RAE avoids this problem completely.

The first thing to note is that, even without adopting the inference condition the adherent of RAE need not say that $B_{10,000}$ is evidence. For what is important here is the process that leads to the beliefs in $B_{1,000} - B_{10,000}$. Consider first the process that leads to believing $B_{1,000}$. This is a process that takes 999 beliefs based on perception as input and then draws an inductive conclusion. The process that leads to believing $B_{1,001}$ is subtly different. It is the process that takes 999 beliefs based on perception, plus one belief based on inductive inference, and draws an inductive conclusion. At each iteration of this process the input beliefs to the inductive inference have a different overall profile. It is plausible to think that as the proportion of the inductively-based input beliefs to perceptually-based input beliefs increases, the reliability of the process goes down. Thus, the reliability of the process that produces the belief B_n ($n > 999$) decreases as n increases. Thus, we do not get an evidential chain reaction as feared.

The second, and more decisive, response to the Urn Objection is to point out that the addition of the inference condition renders RAE immune to the counterexample. Although $B_{1,000}$ is formed via a reliable process, it is an *inferential* process. But the inference condition rules out such processes as the kinds of processes that give evidence. So, RAE + inference condition explicitly maintains that none of $B_{1,000} - B_{10,000}$ are evidence, so long as the process of belief-formation in question is an inferential process.

4.3.3 The Lottery Objection

Imagine that I buy a ticket in a fair lottery with 1 million tickets. In the normal sense of belief, it seems that I will believe that I won't win. Further, even if I don't believe this, there appears to be a reliable belief-forming process available to me that results in the

belief that I won't win. This is the process that begins with my knowledge about the odds of winning the lottery and the fact that I've bought a ticket, concluding with the belief that I won't win. According to RAE, then, it appears that the following proposition is evidence for me:

$\neg W$: I won't win the lottery.

Thus, it should be that $cr(\neg W) = 1$. But, the objection goes, this is absurd. In the situation imagined, it should be that $cr(\neg W) = 1 - (1 \times 10^{-6})$. So RAE gets the wrong result in this case.

The objector can push the point by noting something about how credence is supposed to rationalize betting behavior. If $cr(\neg W) = 1$ as RAE implies it should, then it looks irrational for me to have bought the ticket. But it may not have been irrational. Granting that I have normal sorts of opinions about the value of money, if the ticket cost $1 and if the lottery pays out $2M, then this is a very good purchase indeed. But RAE has the effect of making me look irrational. If $cr(\neg W) = 1$, then as far as I'm concerned, I've suffered a loss of $1 with no chance for gain.

This seems to show that reliability is not a sufficient condition for a proposition being evidence. My belief in $\neg W$ is the result of a very reliable process, and yet it appears that it is not evidence that I have.

Note first that the inference condition may handle this sort of case. If the process that leads to my belief that I won't win is one that is based solely on other evidence that I have, then the inference condition says that $\neg W$ is not evidence for me. However, there is still some reason to pursue a different response. For although it may be the case that $\neg W$ is formed on the basis of inference from one's other evidence, it is not clear to me

that this need be the case. Perhaps after years of training you simply spontaneously form the belief that $\neg W$ whenever buying a lottery ticket. If this is possible then the Lottery Objection still threatens. So, below, I propose a solution to the Lottery Objection that does not depend on the belief in $\neg W$ being a belief arrived at inferentially. This solution will also motivate making a substantive addition to RAE.

In working our way to this addition to RAE, note first that the verdict RAE issues is *not* one that conflicts with something like Lewis's Principal Principle.[24] One might think that it does. In one of its guises, the Principal Principle says that one's credences should be aligned with the known chances. So, the thought goes, it is known that the chance of $\neg W$ is $1 - (1 \times 10^{-6})$. Thus, the $\text{cr}(\neg W)$ should be $1 - (1 \times 10^{-6})$. So, the Principal Principle tells against the advice of RAE. It is important to note that this understanding of the Principal Principle is incorrect. Stated this way, the Principal Principle only applies when there is no inadmissible evidence. For instance, even if you know that the chance that a coin lands heads is ½, you shouldn't set your credence in *Heads* to ½ if you also have been told by a reliable oracle that the coin definitely will land heads. A formulation of the Principal Principle that takes account of this dependence on information takes the following form:

PP: $\text{cr}_G(\bullet) = \text{ch}_G(\bullet)$[25]

Where G stands for the arguments of a chance distribution entailing the chances. The Principal Principle says that if my total evidence is G, then my credences should equal the chances. But if my evidence includes $\neg W$, then the Principal Principle doesn't require my credence in $\neg W$ to equal the chance of $\neg W$. Accordingly, there is no conflict

with the Principal Principle in saying that in the lottery scenario cr($\neg W$) = 1. So, if there is a problem with RAE's handling of the lottery scenario, it is not that it conflicts with the very plausible Principal Principle.

This however, is not a response. It simply notes one thing that the problem is not. There is still the unintuitive feel against the result that RAE issues. One plausible response on behalf of RAE is to note that lotteries induce rather severe context shifts. For instance, it may very well be the case that when one hears a lottery scenario, this has the effect of raising the reliability threshold. Perhaps this can explain why $\neg W$ is not part of the agent's evidence set.

This kind of response is inadequate for several reasons. First, if a response like this were to work across the board, the effect would have to be to raise the standards of reliability to perfection. Otherwise, we could increase the size of the lottery until the inference to $\neg W$ is reliable enough. This is implausible on its own, and it leads to a further problem. When considering lottery cases we seem to think that that I bought a ticket, and what the odds of winning are, are still part of my evidence. But this response with a very high threshold of reliability, would deny that such things are part of my evidence.

The best response, I think, is to add a sensitivity condition to RAE. This amounts to adding a condition that says that for *P* to be evidence, the reliable method indicating *P* must be sensitive. Duncan Pritchard ([2008]) explains sensitivity as follows: "An agent S has a *sensitive* belief in a true contingent proposition $P =_{df}$ in the

---

[24] For Lewis's statement of the Principal Principle see Lewis ([1980]). I state one version of the principle below, in the text.
[25] For this statement of the Principal Principle see Meacham ([2005]).

nearest possible worlds in which *P* is not true, S no longer believes *P*." (p. 438). We can use this to define a sensitive *process* in a similar way:

> A process *p* leading to the belief that *P* is sensitive $=_{df.}$ in the nearest possible worlds in which *P* is not true, *p* does not lead to the belief that *P*.

Accordingly, we add a condition to RAE, insisting that for a reliably indicated *P* to be evidence, it must also be the case that the process is sensitive to the truth of *P*. This does a good job with the lottery case. For in the closest worlds where ¬*W* is false (that is, where the agent *has* won the lottery), the process of belief formation still leads to the belief that ¬*W*. Thus, adding a sensitivity condition permits us to expunge ¬*W* from the agent's evidence set. Of course, if the agent had learned by watching the news that she had failed to win, then ¬*W* still is counted as evidence. For in the closest worlds where ¬*W* is false, *this* method doesn't lead the agent to believe ¬*W* (since she would watch the news, and learn that she had won).[26] The modified version of RAE, then, adds a third condition, which asserts that evidence must not only be such that it is reliably indicated, but reliably indicated by a *sensitive* belief-forming process.

Now, we must be careful with how we formulate the sensitivity condition. For imagine a process that is intuitively *very* sensitive. This process is such that in the vast majority of cases, if *P* is false, the process doesn't result in a belief that *P*. However, according to the definition above, if there is even *one* world in the set of nearest worlds

[26] One alleged problem with sensitivity-based accounts of *knowledge* is that they seem to deprive us of mundane, everyday knowledge. For instance, Sosa ([1999]) asks us to imagine that someone drops a bag of rubbish down the rubbish chute. Several seconds later, does she know that the rubbish is in the basement? Sensitivity accounts of knowledge seem to say 'no', since the closest worlds where the rubbish *isn't* in the chute are worlds where it snagged on the way down. But in such worlds, the agent would still believe it was in the basement. This is thought to be a problem for sensitivity-based accounts of knowledge, since it is allegedly obvious that the agent *does* know that the rubbish is in the basement. However, this does not adversely affect adding such a condition when we're discussing evidence, for

where the process leads to a false belief, then the process is not sensitive. This is surely too stringent: adding such a condition to RAE would require *perfectly* sensitive processes if an agent is to have any evidence.

The thing to say, it seems, is that the sensitivity condition should be stated in a graded way, with probabilities, rather than in the definition above. What is needed is something like:

A process p leading to the belief that P is sensitive to degree $s =_{df.}$

$\text{pr}(\neg(\text{believe } P \wedge p)|\neg P) = s$.[27]

The sensitivity condition then says that for a proposition to be evidence it must be indicated by a process that is sensitive to a high degree. This allows that less-than-maximally sensitive processes ($s < 1$), can still yield evidence. Nevertheless, it rules out processes like the one that leads to the belief that $\neg W$ in the lottery objection.

There is a further advantage to stating the sensitivity condition in this way. As previously mentioned, one could adopt a position weaker than RAE—RCE—which only makes reliability a *necessary* condition on having evidence. One would then be free to add any further conditions to RCE to yield a sufficient condition for having evidence. These further conditions need not have anything to do with reliability. The inference condition adds something very much like this. Now, one might see the

---

there is no intuitive force to the claim that the rubbish being in the basement is *evidence* for the agent in question.

[27] This account of sensitivity is influenced by ideas explored by Sherri Roush in her ([2006]). One might be worried about the function P(•) here. But recall that the existence of such a function is a presupposition of reliabilist epistemologies. Thus, I simply take this for granted. If there is no such function, then reliabilist epistemology, in general, is in a lot of trouble, never mind my specific application of it. Note also that some kind of counterfactual reasoning is still going on in evaluating the probability appealed to here. In particular, we need to use some notion of closeness to evaluate the value for the probability function, conditional on $\neg P$ and p. Using probability simply lets us appeal to a graded notion of sensitivity. This appeal to closeness is no different than the appeal to some kind of closeness in evaluating the reliability of a process that features in condition (ii) of RAE.

addition of a sensitivity condition to RAE in the same way: I am advocating reliability as a necessary condition, and then adding the sensitivity condition to yield an account that gives necessary and sufficient conditions for having evidence. However, adding the sensitivity condition to RAE is importantly *different* than adding the inference condition. Both, I believe, are motivated additions, however unlike the inference condition, the sensitivity condition doesn't itself smuggle in extra non-reliability considerations.

Note first that as we've been talking about the reliability of a process *p*, we've understood it in the following way:

$$\text{pr}(P|\text{b}P \wedge p) = r$$

where *r* is the reliability of the process *p*, and 'b*P*' abbreviates 'believe *P*'. This says that the probability that *P* is true, given that you believe *P* using the process *p* is *r*. In frequency terms, this says that the proportion of true beliefs yielded by *p* is *r*. This is certainly the standard way to understand such reliability claims. Now, the sensitivity condition requires a high value for $s = \text{pr}(\neg(\text{b}P \wedge p)|\neg P)$. As we'll see, *ceteris paribus*, a high *s*-value for this boosts the reliability value, *r*. So all that requiring a high *s* value adds to RAE is a requirement that there is a certain *way* in which processes must be reliable.

According to Bayes' Theorem we have:

$$\text{pr}(P|\text{b}P \wedge p) = \frac{\text{pr}(\text{b}P \wedge p|P) \times \text{pr}(P)}{[\text{pr}(\text{b}P \wedge p|P) \times \text{pr}(P)] + [\text{pr}(\text{b}P \wedge p|\neg P) \times \text{pr}(\neg P)]}$$

Note that $\text{pr}(\text{b}P \wedge p|\neg P) = 1 - s$, given that $\text{pr}(\neg(\text{b}P \wedge p)|\neg P) = s$. Thus, as the sensitivity of process *p* increases, the value of $\text{pr}(\text{b}P \wedge p|\neg P)$ decreases. Note, however, that as this happens, $\text{pr}(P|\text{b}P \wedge p)$ increases. So, requiring high sensitivity does not allow *less*

reliable processes of belief formation to yield evidence propositions. Rather, what requiring high sensitivity does, is rule out certain situations where pr(P|b$P \wedge p$) is high purely on account of a high prior probability for P. For instance, if the prior probability for P is extremely high, say 0.999, then a high value of pr(b$P \wedge p$|$\neg P$) will have little effect on the value of pr(P|b$P \wedge p$) since pr($\neg P$) is so small. This is exactly what is happening in the lottery scenario where the prior probability that you will lose is so immensely high that a non-sensitive process of belief formation can still have a high value for pr(P|b$P \wedge p$), and thus count as reliable. Adding the sensitivity condition does rule this kind of reliable belief-forming process as giving evidence, but it does it in a way that is based purely in reliability considerations. Thus, the sensitivity condition is fully in the spirit of RAE, and does not smuggle non-reliability considerations into the account.

In light of this, we can formulate a more perspicuous statement of RAE as it would look with the sensitivity condition:

> **RAE + sensitivity condition:** The set of propositions $\mathbf{E_t}$ is S's evidence set at t iff
>
> (i) for each $P_i \in \mathbf{E_t}$ there are reliable belief-forming processes, $\mathbf{p}_i$ available to S at t such that if S applied those operations S would believe all the P_i at t and those beliefs would be caused by reliable belief-forming processes,
>
> (ii) pr($\neg$(b$P_i \wedge \mathbf{p_i}$)|$\neg P_i$) $\geq$ s.

The account that I want to defend, however, incorporates both the inference condition *and* the sensitivity condition. Thus, the final reliability-based account of evidence, which I think is worth defending is the following:

> **RAE*:** The set of propositions $\mathbf{E_t}$ is S's evidence set at t iff

(i) for each $P_i \in \mathbf{E_t}$ there are reliable belief-forming processes, $\mathbf{p}_i$ available to S at t such that if S applied those operations S would believe all the P_i at t and those beliefs would be caused by reliable belief-forming processes,

(ii) none of the $\mathbf{p_i}$ are inductive inference from S's other beliefs, and

(iii) $pr(\neg(bP_i \wedge \mathbf{p_i})|\neg P_i) \geq s$.[28]

As a side comment it is worth noting that the sensitivity condition may permit a different response to the Urn Objection. For instance, to see if $B_{1,000}$ is evidence we must go to worlds where $B_{1,000}$ is false, and see if the method in question leads the agent to believe $B_{1,000}$. If the closest worlds where $B_{1,000}$ is false are worlds where the first 999 draws are black, then the inductive process of belief formation will yield the belief that $B_{1,000}$ and so the process will not be appropriately sensitive. Thus, $B_{1,000}$ is not evidence, the chain reaction is blocked on the first step, and we have a response to the Urn Objection. However, if the set of closest worlds where $B_{1,000}$ is false includes worlds where the draws are different, then it is likely that the process in question would not lead to the belief that $B_{1,000}$. If that's the right way to think about this case, then the process is sensitive. Since it is not clear exactly what the closest worlds are like, it is unclear whether or not this sensitivity condition can provide a response to the Urn Objection.[29]

[28] If our starting model is RAE-t rather than RAE, we would add a fourth condition: (iv) every member of $\mathbf{E_t}$ is true.

[29] Note that facts about how the process is specified may end up fixing facts about the closest worlds. For instance, if the relevant process is: forming the belief $B_{1,000}$ in response to the truth of B_1 - B_{999}, then there will be no worlds where *that* process is used, and where B_1 - B_{999} are not all true.

4.3.4 The Scientific Instrument Objection

Consider a scientific setting where there is an instrument that has a small and known error rate. Let's say that the instrument is a detector of β-particles, and it is well known that this detector delivers accurate readings 99% of the time.[30] This would seem to be reliable enough to make what it says about the presence of β-particles evidence. That is, when the detector says that there is a β-particle, the proposition that there is a β-particle is evidence. But then such propositions should be assigned credence 1. But, it seems, this is absurd. According to RAE, cr(β-particle) = 1, and yet this is a prime situation in which it should be that cr(β-particle) = 0.99.[31]

Note that the sensitivity condition doesn't help here. For if the detector works as I have described, then it is *extremely* unlikely that the detector will say there is a β-particle given that there is no β-particle.[32] Thus, the value of *s* is very high.

The inference condition, as stated in RAE*, might give one a response to this sort of scenario. For, the normal way of understanding this case is one where the scientist sees that the detector says there is a β-particle, and then infers that there *is* a β-particle. But if this is the process that leads to the belief that there is a β-particle, then that there is a β-particle is not evidence. This response, however, depends on it being true that the belief that there is a β-particle results from such an inference. The clearest

---

[30] Though the details don't matter for the example, we can make this precise by stipulating that the detector never misses a β-particle (so we have no false negatives), but that 1% of the time it indicates that there is a β-particle, there is not.

[31] In some ways this is just a precise way of putting the objection considered in Chapter 3 concerning less than maximally reliable process yielding full credence.

[32] Let "β" = the detector says there is a β-particle; β = there is a β-particle. Above, I said that p("β"|β) = 1, which implies that p(¬"β" ∧ β) = 0. If we let p("β" ∧ β) = *n*, then from the fact that p(β|"β") = 0.99, we can conclude that p("β") = *n*/0.99. Thus, it must be that p("β" ∧ ¬β) = (0.0101… × *n*). Further, since p(β)

way of understanding this scenario is in this way. Thus, RAE* can take care of the clear counterexample. However, just as in the Lottery Objection, this may not be the only way to understand the scenario. Perhaps we can tell the story in such a way that the belief that there is a β-particle is very reliable, *not* the product of inference, and yet we still feel that it should not receive full credence. If there were such a case, RAE* would be vulnerable. I am not certain there is such a case, but I am also not certain there is not. Accordingly, in what follows I will show how RAE* can *still* respond to such an objection.

The response to the Scientific Instrument Objection is to point out that there is a case-by-case fix that can be adopted in such scenarios. Consider the visual process leading to the belief that the detector says there is a β-particle ("β"). Grant that this process is reliable to degree 0.95, which means that in 95% of the worlds in which the scientist form a belief in this way, the belief is true. Now consider the process that leads to the belief that there is a β-particle (β). We are not imagining this belief as being *inferred* from the belief that "β" (perhaps the scientist doesn't even have such a belief), however, it seems clear that in the scenario the reliability of the process leading to the belief that β is somehow dependent on the scientist's ability to tell when the detector says that there is a β-particle. If that's the case, then the process, based on the detector reading, leading to the belief that there *is* a β-particle (β) is less reliable than the process leading to the belief that "β". This is because the process leading to the belief that β inherits the unreliability of the scientist's visual process and *adds* to it the unreliability of the detector. By setting the level of reliability needed for a process to yield evidence

---

= p(¬"β" ∧ β) + p("β" ∧ β), we know that p(β) = *n*, and so p(¬β) = 1 – *n*. Thus, p("β"|¬β) = (0.0101… ×

somewhere between the reliability of the process that leads to the belief that "β" and the process that leads to β, we can guarantee that "β" is evidence while β is not. As long as the scientist has well-calibrated conditional credences (e.g., cr(β|"β") = 0.99), conditionalizing on "β" will then yield cr(β) = 0.99 as desired. For all such situations, there will be some level of reliability such that if we set the required level for evidence at this level, we will get the desired result.

One might reasonably ask *why* we are to set the reliability at just this level. In response to this, we can note that the level of reliability can be set at different places for different evaluative purposes. All I point out is that there *is* a level of reliability such that when we set the reliability necessary for evidence at that level, we get the intuitively correct result. In fact, it seems wrong to say that the intuitively appropriate credence in β will *always* be 0.99. If we're not interested in the details of the scientific instrumentation, it might be appropriate to gloss β as evidence and thus assign credence 1 to it. Of course, in the objection, the case was described in such a way to make us think that the object of inquiry for the scientist is the presence or absence of β-particles. But in such a situation we have a nice explanation for why we won't want to take the presence of a β-particle as evidence.

4.3.5 Summary

RAE* is a strong thesis: it gives necessary and sufficient conditions for a proposition being in an agent's evidence set. Given this, it generates many potential counterexamples. I have attempted to defend RAE* against the most striking of these

---

n)/(1 – n). If n = 0.1, then p("β"|¬β) ≈ 0.0011. Even if n = 0.5, p("β"|¬β) ≈ 0.01.

counterexamples. I believe that RAE* is adequately defended against such counterexamples. But some might not be so sympathetic.

Note that all the objections in Section 3 have aimed at attacking reliability as a sufficient condition for some proposition being a member of an agent's evidence set. Thus, even if one is unsympathetic with my defense of RAE against these objections, one could still be open to taking reliability as a *necessary* condition on a proposition being in an agent's evidence set. In the next section, then, I formulate RAE as a necessary condition and show that it can still do interesting work.

4.4 The Reliabilist Constraint on Evidence

The amended proposal is:

> **Reliabilist Constraint on Evidence (RCE):** The set of propositions $\mathbf{E_t}$ is S's evidence set at t only if for each $P_i \in \mathbf{E_t}$ there is a process $\mathbf{p}_i$ such that $\text{pr}(P_i|\text{b}P_i \wedge \mathbf{p}_i) \geq r$ (where all the $\mathbf{p}_i$ are available to S at t).[33]

Note that as a necessary condition, RCE does say that there is a lower bound on the reliability of the processes that give one evidence. Further, if one wants some proposition to be evidence that is indicated with reliability to degree (say) 0.8, then one must be open to other propositions indicated with this degree of reliability to be evidence. RCE doesn't *mandate* that every such proposition is evidence (since a further

[33] I do not state the sensitivity condition or the inference condition, since they were added to respond to objections to RAE which do not apply if we adopt RCE which is only a necessary condition on being evidence. Of course, one might find it useful to formulate these conditions as a *additional* necessary condition on being evidence. Further, I do not include a truth condition in RCE, though one could.

necessary condition on being evidence could rule them out), but it does require one to say something substantive to rule them out. So, RCE is a genuine constraint.[34]

RCE can be tacked on to many different kinds of accounts of evidence. In fact, RCE is compatible with internalist and externalist accounts of evidence. One could, for instance, pair MSE with RCE. Roughly, such an account would say that one's evidence consists of the facts about one's mental state to which one has a reliable route. Nevertheless, RCE does rule out certain views. It is, for example, not compatible with an account of evidence according to which you must have access to whether or not something is part of your evidence. Since one does not often have access to whether or not one's beliefs are reliably formed, one does not have access to whether or not a certain belief is a piece of evidence.

4.5 Attractive Features of RAE/RCE

4.5.1 Answering Neta's Question

In the latter half of his ([2008]), Neta considers belief-independent accounts of evidence (he calls such accounts 'non-doxastic accounts'). After rejecting some initial ways of formulating such an account, he proposes two accounts that he finds attractive:

> **(ND1)** *P* is a member of S's evidence set at *t* if and only if, at *t*, S has a perceptual experience or apparent memory that has the representational content that *P*.

---

[34] More precisely, whereas RAE gave us:

> (NEC) If *P* is not evidence (and *P* is reliably indicated at level *r*), then there are no evidence propositions indicated at a level < *r*.

RCE would give us:

(ND2) *P* is a member of S's evidence set at *t* if and only if, *P* is a true proposition to the effect that, at *t*, S has a particular perceptual experience or apparent memory.

Although Neta is attracted to both these accounts, he worries that neither proposal is well-motivated. What is it, he asks, about experience and memories that make them so evidentially special?

> We should need to hear a good answer to this question in order to be moved to accept either of these two non-doxastic account of evidence. Until I hear such an answer, I leave it open that either account is correct, but I regard each as insufficiently well motivated: they leave us in the dark as to why it is that just these mental states determine what's in our evidence set, and so determine how we should distribute our confidence in hypotheses. (p. 111).

If RAE* is correct, then I think we can give an answer to Neta's question. Why is it that (ND1) and (ND2) strike us as initially plausible accounts of an agent's evidence? Perhaps it is because we are usually highly reliable about the content of our memories and experiences are, and these memories and experiences are often highly reliable guides to the way that things were and are. There is, then, nothing evidentially special about memories or experiences just in virtue of the fact that they are memories or experiences. However, there is an important contingent fact: humans tend to be very reliable about the content of their own memories and experiences. If RAE* is true, this explains the appeal of accounts like (ND1) and (ND2).

I think that the ability of RAE* to explain intuitive judgments like this should not be underestimated. In particular, by appealing to reliability considerations, RAE* offers us the prospect of an epistemologically satisfying and genuinely explanatory account of what it is to have evidence. This stands in stark contrast to an account of

---

(NEC) If *P* is not evidence and (and *P* is reliably indicated at level *r*) *and P meets the other*

evidence according to which evidence is the output of certain faculties. Without an account like RAE*, such accounts do look unmotivated, as Neta points out.

Now, if RCE is the account we go with, then we don't get quite the same explanation, for reliability alone doesn't determine what evidence an agent has. But according to RCE, reliability *is* a genuine constraint on evidence, and we can use this to at least partially answer Neta's question in the way indicated above. This underlines a general advantage of RCE: it promises also to partially explain what it is to have evidence.

4.5.2 Handling Clear Cases

4.5.2.1 Pete and Tom

Consider the example from Chapter 2. Imagine that we have two ordinary humans, Tom and Pete, who are just about to walk out of their air-conditioned building in New York City. At t_0, just before going through the door Tom's credence function is such that cr_0(It is hot in NYC today) = 0.5. Then, at t_1, after walking outside Tom's credence in this proposition goes to 1, with the rest of his beliefs being properly conditionalized on this information. On the other hand, at t_0 Pete's credence function is such that cr_0 (It is hot in Syndey today) = 0.5. Then, at t_1, Pete's credence in this proposition goes to 1, with the rest of his beliefs being properly conditionalized on this information. *Prima facie*, Tom's change of belief is more rational than Pete's. Both RAE* and RCE deliver this verdict. For, remembering that Tom is a normal human, Tom's belief that it is hot in NYC is produced by a reliable belief-forming process, namely that he feels the heat

---

conditions on being evidence, then there are no evidence propositions indicated at a level < *r*.

and then concludes that it is hot where he is (which he knows to be NYC). The belief that it is hot in Sydney today, however, is not formed by such a reliable belief-forming process. Thus, it is not evidence, and so Pete shouldn't have conditionalized on it.

4.5.2.2 Funny Conditionalization

RAE* and RCE can also both handle the case of inappropriate conditionalization from Chapter 2. Recall that the problem there was that there are two ways consistent with updating via conditionalization. First, one gets evidence *E*, and updates *H* on this. Second, one decides to believe *H*, and so believes *E* in such a way to make it look as if one conditionalized on it. RCE says that the second scenario is in violation of conditionalization, since the way in which the agent comes to believe *E* in this situation is certainly not a reliable belief-forming process. Believing that there is evidence to justify one's beliefs is not a reliable belief forming process. Accordingly, it rules out this way of updating.

It is worth noting that something like MSE cannot naturally handle this case. For we could stipulate that *E* is entailed by the agent's current mental state. Thus, *E* is evidence according to MSE. Nevertheless, it is still possible for the agent to update in these two different ways. That is, it is possible for the agent to believe *E* because there is a reliable process leading to this belief, or it is possible for the agent to believe *E* because he believes *H* and this evidence would support *H*. These are both possible even if *E* is entailed by the agent's current mental state. If we restrict the term 'available' in RCE to those processes actually used, then RCE can distinguish these cases in a way that MSE cannot. So, if MSE is to handle such cases, RCE should be adopted as an extra constraint.

4.5.2.3 Baseball on the Head

Consider a different case. Jessica is walking along next to the baseball field and suddenly gets hit on the head by an errant fly ball. Suppose that this results in Jessica firmly believing *R*: it is raining. Suppose that it has, in fact, just started to rain. RCE says that *R* is not evidence for Jessica, even though it is true and she firmly believes it, because the process of forming beliefs as a result of strikes on the head is not a reliable process. Alter the case slightly, so that Jessica's strike on the head makes it appear as though it is raining. Even then, RCE says that *R* is not evidence for Jessica, for the process is no more reliable just because it appears as if it is raining to Jessica. On the other hand, consider the proposition *AR*: it appears as if it is raining. If Jessica has a reliable process of forming beliefs about how things appear to her (and this is not affected by the strike on the head), then *AR* will be evidence for Jessica in this situation. This is an intuitively satisfying result, and seems to provide an appropriate way to epistemically evaluate agents.

4.5.3 Picking Appropriate Evidence Propositions

There is one final benefit of RAE*/RCE that is worth mentioning. RAE*/RCE is helpful in picking out the appropriate propositions that constitute one's evidence. Something like MSE says that one's evidence consists of those things consistent with one's entire mental state. But though this has some initial plausibility, it doesn't ultimately seem correct to say that every feature of one's mental state is part of one's evidence. RAE*/RCE puts an extra constraint on evidence. It says that only those aspects of one's mental state to which one has reliable route available are part of one's evidence. Consider how this goes in a case of visual perception. Although there might

be a mental representation of the scene before one with a high level of detail, it doesn't seem right to say that *every* feature of the scene that is represented is part of one's evidence. Rather, what is part of one's evidence are those parts of the scene, which one can reliably extract. RAE*/RCE picks out propositions describing these features as the agent's evidence

Most of the examples we have considered have been perceptual examples like this. We wonder what the agent's evidence is *now*, given some perceptual processes. But, intuitively, one also has evidence from memory. RAE*/RCE seems to do well in this arena, too. The exact way in which evidence from memory would work according to RAE*/RCE is dependent on a more fully specified picture of how an agent's memory system is set up. But consider the following rough sketch: Our memory system is a system that both allows the explicit formation of conscious beliefs (so it is like a perceptual system) and also works as a belief-maintenance system that maintains implicitly held beliefs that may or may not be conscious.

Consider first the implicitly maintained beliefs. According to RAE*, those implicit unconscious beliefs that are reliably maintained are evidence. Some plausible candidates here are general beliefs about one's environment, one's name, and one's occupation. It seems just right to say that, in normal situations, these propositions are evidence.

Consider now those beliefs that are explicitly and consciously formed as a result of "searching" through one's memory. RAE* says that these beliefs are evidence just so long as they are reliably formed.[35] Sometimes belief in the thing remembered will be

[35] And the other two conditions are met.

reliable and so the thing remembered will be evidence. But more often, such beliefs will not be produced in a reliable enough way, and so will not be evidence. Rather, what will be evidence is something about what one seems to remember, or perhaps the thing remembered, but with many of the specific details left out. Instead of having a reliable route to what one had for lunch in Barcelona four years ago, one merely has a reliable route to the fact that one was in Barcelona some years ago. On reflection, these kinds of propositions seem to be just the kind of thing that are evidence. RAE*/RCE, it seems, is able to select in an intuitively satisfying way, those propositions that are evidence from those that are not.

4.6 Conclusion

In Chapter 3 I formulated a specific account of evidence, RAE, which is designed to adequately work with existing Bayesian Epistemology. In this chapter I set myself the task of responding to objections, both general objections to EE, and more specific ones to RAE. After considering the objections, I made some substantive additions to RAE: the inference condition and the sensitivity condition. The resulting view, RAE*, is a promising reliability-based account of having evidence. I concluded by showing that a weaker, necessary condition on evidence (RCE) can still do work in giving an account of evidence. This is of interest, because even if one is not attracted to RAE*, one could still endorse RCE.

In Chapter 5, I'll turn away from a general assessment of RAE*, and focus instead on an especially puzzling feature of evidence, and show how RAE* can handle such a phenomenon in an attractive way.

CHAPTER 5

LOSING EVIDENCE

5.1 Introduction

In the last two chapters, I have investigated several ways of understanding what it is to have evidence, all of which are based on reliability considerations. However, I have allowed one important ambiguity in the stating each of these views: I have not said whether the accounts give an agent's *total* evidence at *t*, or whether they give the agent's *new* evidence at *t*. This issue will be important in this chapter.

For the moment, focus on RAE*:

RAE*: The set of propositions $\mathbf{E_t}$ is S's evidence set at *t* iff

(i) for each $P_i \in \mathbf{E_t}$ there is a process $\mathbf{p}_i$ such that $\text{pr}(P_i|\text{b}P_i \wedge \mathbf{p}_i) \geq \text{r}$ (where all the $\mathbf{p}_i$ are available to S at *t*),

(ii) none of the $\mathbf{p_i}$ are inductive inference from S's other beliefs, and

(iii) $\text{pr}(\neg(\text{b}P_i \wedge \mathbf{p_i})|\neg P_i) \geq \text{s}$.

If we take this as giving an account of the agent's *new* evidence at *t*, then the natural picture is one where old evidence is simply maintained by fiat. New evidence is conjoined with the old evidence, so that one's total evidence either increases or stays the same as time passes. This picture could be useful for certain purposes. One shortcoming of such an approach, however, is that it does not allow evidence to be *lost*.

If we understand RAE* it as giving an account of the agent's *total* evidence at *t*, then we open up the possibility of lost evidence. The fact that *P* was evidence yesterday, does not entail that *P* is still evidence now. For *P* to be evidence now, there must *now* be a reliable belief forming process indicating *P*. This is a more realistic

picture in some ways. For one, it allows us to say plausible things about forgetting. Imagine that twenty years ago I knew that my gym locker combination was *5-16-22*. Perhaps this was even evidence for me twenty years ago. Nevertheless, if I have now forgotten that my locker combination was *5-16-22*, it doesn't seem appropriate to say that fact is still evidence for me now. A plausible story here is that that fact is not now evidence, because I no longer have a reliable route to its truth. If RAE* gives an account of the agent's *total* evidence at *t*, then it can account for these kind of situations.

I think we'd like to have an account of evidence that allows for lost evidence. I just sketched how RAE* can handle lost evidence due to memory failures in a plausible way. Note, however, an assumption that we must make if RAE* is to be able to account for lost evidence. We must insist that the initial process of belief formation (which made *E* evidence) is distinct from the process that maintains this belief at later times. For if we *don't* grant this, then the process that indicates *E* is just the initial and highly reliable process that led to *E* being evidence in the first place. RAE* would thus say that *E* is always evidence. Thus, we must say that the process relevant to the belief that *E* at one time is distinct from the process relevant to the belief that *E* at a later time. Exactly how to understand this is tricky, but the basic idea can be pictured as follows (where t_0, t_1, t_2 are subsequent epistemic moments):

t_0: Belief that *E*. Process: normal visual processing

t_1: Belief that *E*. Process: normal visual processing + maintenance

t_2: Belief that *E*. Process: normal visual processing + more maintenance

In this chapter, I would like to investigate a way of losing evidence that is different from the way we lose evidence in a case of memory loss. In particular, I will

be interested in cases where an agent loses evidence because he comes to get more evidence that tells against the initial evidence. I will explain how RAE* can model at least a restricted class of such cases of evidence loss. In the account, however, it will be important to keep in mind the picture of belief-formation that is in the background. It is one according to which there are distinct processes of belief formation for belief in the same proposition at different times. I have noted above why we need this picture if we are to explain the loss of evidence in mundane cases involving forgetfulness. In what follows I will help myself to this basic metaphysical picture. I will not be defending this picture, so for the purposes of this chapter, it is an assumption. If it is an assumption that is radically mistaken, how RAE* handles evidence loss would have to be reworked.

5.2 The Phenomenon of Undercut Evidence

I take it as a starting point that agents can lose evidence in various ways. One common way of losing evidence, discussed above, is through forgetfulness or memory loss. But there is another way in which it is plausible that an agent can lose evidence. Consider the following story:

> I am looking out my window and see the snow falling. It is plausible that *S* = "It is snowing," is a member of my evidence set. But now imagine that I am told that the tenants in the floor above me are playing a practical joke, and dropping fake snow out of their window. Upon learning this, it seems that *S* should no longer be a member of my evidence set.[1]

Consider a different example:

[1] Feldman ([1988]) discusses a case similar to this one.

> It appears to me as if it is snowing. At the very least, it would seem that *AS* = "It appears as if it is snowing," is a member of my evidence set. But now imagine that a neuroscientist tells me that I'd believe it appeared to be snowing even if it didn't. Upon learning this, it seems that *AS* should no longer be a member of my evidence set.

Both of these cases are examples where an agent's evidence appears to be undercut or defeated by learning some new information.

It will help, in what follows, to deal with a particularly "pure" case of undercut evidence. So, consider the Red Jellybean Case[2]:

> Wally walks into a room and sees a red jellybean on the table. The proposition that the jellybean is red is plausibly a member of Wally's evidence set. Wally is then told that the room is lit entirely with red lights. It is plausible that the proposition that Wally is told this subsequently becomes a member of Wally's evidence set. In such a situation, it seems as though the proposition that the jellybean is red should no longer should be a member of Wally's evidence set.

Bayesian accounts struggle to deal with evidence loss in this kind of way. Reliabilism struggles to deal with the phenomenon of epistemic defeat. Thus, it is somewhat surprising that RAE* together can combine with a Bayesian account to handle cases like this in an attractive way. Nevertheless, I will attempt to show that this is the case.

[2] This follows closely a scenario discussed in Weisberg ([2009]).

5.3 The Problem of Undercut Evidence

5.3.1 First Problem: Conditionalization

Undercut evidence is a problem for Bayesian accounts for two reasons. The first reason is fairly clear. According to standard Bayesian accounts, one's belief state is updated by conditionalizing on one's evidence. Conditionalization is usually understood as a sequential updating rule. Assume that time can be divided into discrete epistemic instants, moments at which epistemic changes happen. According to the standard understanding of conditionalization, at each instant *t*, you update your credence function at $t - 1$ on the new evidence received at *t*. That is:

(sq-*COND*) $\mathrm{cr}_t(\bullet) = \mathrm{cr}_{t-1}(\bullet|E_t)$

where E_t is the new evidence received at *t* (or: between $t - 1$ and *t*). Since $\mathrm{cr}_{t-1}(E_t|E_t) = 1$, it follows that $\mathrm{cr}_t(E_t) = 1$. Given this, $\mathrm{cr}_t(E_t|\bullet) = 1$, and since cr_t will be changed only by conditionalization, $\mathrm{cr}_{i>t}(E_t) = 1$. Thus, evidence once received is kept forever. This, however, is a problem if we think that evidence can be lost, whether by way of forgetfulness or by being undercut by further information.

5.3.2 Second Problem: Rigidity

The second reason that undercut evidence is a problem for Bayesian accounts is less obvious, but was brought to light recently by Jonathan Weisberg ([2009]). To see the problem, we must first say something about the phenomenon of epistemic defeat. In general, a defeater is something (usually a proposition) that in some way removes the epistemic support that some proposition once enjoyed. Pollock & Cruz ([1999]) characterize defeat as follows:

> If *E* is a reason for *S* to believe *H*, *D* is a defeater for this reason if and only if *D* is logically consistent with *E*, and *E&D* is not a reason for *S* to believe *H*. (p. 37)

Though I will not use this as a definition of 'defeater' it does help to focus on the concept in question. Defeaters are generally classified into two categories: rebutting and undercutting defeaters. A rebutting defeater is a defeater that provides positive reason to believe the negation of the hypothesis in question (¬*H*). An undercutting defeater is a defeater that doesn't provide positive reason to believe ¬*H*, but instead neutralizes whatever positive support *H* enjoyed.

We are interested in cases where some proposition in one's evidence set is defeated. Often, when one's evidence is defeated it is a form of undercutting defeat. The Red Jellybean example at the beginning of this paper is a plausible instance of undercutting defeat. In this case, the learned information (*TRL*) doesn't render the evidence proposition (*R*) unlikely on its own, but only serves to nullify the support that the evidence proposition previously received. Evidence propositions obviously can also be subject to rebutting defeat, but it is undercutting defeat which raises particularly difficult puzzles.

Consider how we might model such undercutting defeat in Bayesian terms. Let *H* be the target proposition that is believed by the agent, but will be undercut. (If we are thinking about undercut *evidence*, then *H* will be a proposition in the agent's evidence set). Let *D* be the proposition that is a defeater, and *E* the reason that supports *H*. Assume further that *E* is not only the reason that supports *H*, but further that *E* is a proposition in the range of the agent's credence function. Given this, we can say that *E* is a *doxastic* reason for believing *H*. It may help to think of an example. Let *H* = "the

Red Sox won last night", E = "espn.com reported the Red Sox won last night", and D = "espn.com is experiencing problems and has not been updated for 24 hours."

Initially, D is irrelevant to H. So, $cr_0(H) = cr_0(H|D)$. After E provides a reason to believe H, however, D is relevant to H. So, $cr_1(H|E) < cr_1(H|D \wedge E)$. This fits well with a natural way to think about undercutting defeaters:

D is an undercutting defeater for H (with respect to E) just in case

(I) $cr(H|D) = cr(H)$

(II) $cr(H|D \wedge E) < cr(H|E)$[3]

But now consider a case where the reason in favor of H is not a doxastic reason, like E, but rather a non-doxastic reason. I'll understand a non-doxastic reason for H as being something that gives an epistemic reason to believe H and is not itself in the range of the agent's credence function.[4] A paradigmatic case of a non-doxastic reason for belief comes in cases of evidence acquisition. A certain experience may be a non-doxastic reason to believe that H, or being in a certain mental state may be a non-doxastic reason to believe that H. Alternatively, facts about reliability may furnish non-doxastic reasons to believe H. The Red Jellybean Case gives an example of this. Let H = "the jellybean is red", and let D = "the room is illuminated by red lights." In this case, there is no proposition E that supports H. Instead, H receives non-doxastic support, from the experience that Wally has or from the fact that the process leading to the belief

[3] See Kotzen, ([*ms*]), who attempts to give a formal account of epistemic defeat. He starts out with something very similar to (I) and (II) (see p. 8), and then modifies the account to get around problems unrelated to the focus of this paper.

[4] Note that there are two ways to be a non-doxastic reason to believe H. First, something could be a non-doxastic reason for H if that thing is a proposition that supports H, but is not in the range of the credence function. Second, something could be a non-doxastic reason for H if that thing is not a proposition but it supports H. Since the thing in question is not a proposition, it cannot be in the range of any credence function.

that H is reliable. Nevertheless, it seems that D is still an undercutting defeater for belief in H.

But consider how the scenario goes. Initially, D is irrelevant to H. So, $cr_0(H) = cr_0(H|D)$. Then, after Wally looks at the jellybean, he has a non-doxastic reason to believe H, and D is relevant to H. So, $cr_1(H) < cr_1(H|D)$. Note several things about this. First, from *t0* to *t1*, the credence assigned to D need not change. What changes is the relevance that D has to H. Second, notice that since the reason to believe H was a non-doxastic reason the only proposition in the agent's credence function that is directly affected from *t0* to *t1* is to H.

From this, it follows that a condition called *Rigidity* was violated in the credal change from *t0* to *t1*. Rigidity says that when the only directly affected propositions between *t0* and *t1* are those in the partition $\{E_i\}$ $cr_1(\bullet|E_i) = cr_2(\bullet|E_i)$. Applied to the Red Jellybean Case, this entails that $cr_0(D|H) = cr_1(D|H)$. Recall that $cr_0(H) = cr_0(H|D)$. From this it follows that $cr_0(D) = cr_0(D|H)$.[5] Together with Rigidity, this entails that $cr_1(D|H) = cr_0(D)$. Since it was part of our story that there was no change to the credence assigned to D from *t0* to *t1*, $cr_0(D) = cr_1(D)$. From this it follows that $cr_1(D|H) = cr_1(D)$. But this means that $cr_1(H|D) = cr_1(H)$. But if D is an undercutting defeater, then $cr_1(H) < cr_1(H|D)$. So Rigidity is violated in this scenario. However, Conditionalization implies Rigidity, so Conditionalization is violated in this scenario. So, having a proposition that enjoys non-doxastic support undercut conflicts with Conditionalization. Since evidence propositions enjoy non-doxastic support, there is a conflict between Conditionalization

[5] See p.181, footnote22.

and undercut evidence. This is the second problem for giving a Bayesian account of undercut.

Note that the two problems are distinct. Even if we solved the first problem (which concerns the fact that certainties are maintained forever) we'd still be faced with the problem concerning Rigidity. As Weisberg [2009] writes:

> …the basic problem is that a probabilistic correlation between [*D*] and [*H*] needs to be introduced when the experience *E* is had. Initially, [*D*] has no probabilistic bearing on [*H*], but it should have a negative bearing on [*H*] once [*H*] has been boosted on the basis of *E*. Rigidity, however, prevents any such correlation from being introduced when *E* has its effects. (pp. 14-15).

The root of the problem is that an undercutting defeater is one that does not tell against the target proposition until the reason for the target proposition is active. We can model this well in Bayesian terms when the reason is a doxastic reason. But when the reason is a non-doxastic reason, this account fails. So the notion of non-doxastic undercutting defeat is one that cannot be modeled on standard Bayesian accounts.

5.4 Hypothetical Prior Conditionalization

In this section, I will discuss the first problem. We need to understand conditionalization in a way different than sq-*COND* if there is any hope of handling cases of undercut evidence.

One might think that the appropriate response is to move away from conditionalization and towards Jeffrey Conditionalization, where evidence need not receive credence 1. According to Jeffrey Condtionalization:

(j-*COND*) $\mathrm{cr}_t(\bullet) = \sum_i \mathrm{cr}(\bullet|E_i) \times \mathrm{cr}(E_i)$

where the E_i together form a partition over propositions representing the agent's evidence. Although I have sympathy with this suggestion, Jeffrey Conditionalization is easily trivialized. We need to say something substantive about the input partitions to Jeffrey Conditionalization if we are to get a helpful epistemological account.[6] In Chapter 3 I discussed the difficult problems facing such a project, which I will not repeat here. Even if these problems could be overcome, an account of evidence that fits with Jeffrey Conditionalization would face grave difficulties of the sort described in Christensen ([1992]). Thus, I stick with standard conditionalization, according to which one's evidence receives credence 1. Something must be said, then, to fix conditionalization.

We can fix conditionalization by moving away from sq-*COND* and towards what we can call 'hypothetical prior conditionalization' (hp-*COND*).[7] According to hp-*COND*, at each epistemic moment *t* you update your hypothetical prior function on your total evidence at *t*. Let 'hp' refer to the agent's hypothetical prior function. Then:

(hp-*COND*) $\mathrm{cr}_t(\bullet) = \mathrm{hp}(\bullet|E_t)$

where E_t is the agent's *total* evidence at *t*. hp-*COND* is slightly different than sq-*COND* because it allows that an agent can lose evidence. Say that at t_1, E_{t1} is reliably indicated to agent S. It can happen that at some later time t_2, E_{t1} is no longer evidence for S. Perhaps this is because S has a very bad memory. More interestingly, perhaps this is because S has received new information that leads him to have strong skeptical background beliefs. If something like this happens, then E_{t1} may not be part of S's evidence set at t_2. Accordingly, $\mathrm{cr}_{t2}(E_{t1})$ need not be 1, since $\mathrm{hp}(E_{t1}|E_{t2})$ need not be 1.

[6] See Weisberg, ([*forthcoming*]) "Varieties of Bayesianism."

If we understand conditionalization in this way, then it is possible for agents to lose evidence. At every moment, we can picture the set of worlds to which some agent, S, assigns positive credence. According to sq-*COND*, this set of worlds either stays the same size or contracts as time passes. According to hp-*COND*, this set of worlds can stay the same, contract, *or expand* as time passes.

This, then, answers the first worry concerning conditionalization and everlasting certainties. It makes it possible to give a satisfying account of how evidence can be undercut by new information. Of course, it is only a first step in this direction. To complete the story, one needs to say something about how propositions can be removed from an agent's evidence set. In a moment, I will attempt to explain how this can be done. But first it is important to clear up some things about hp-*COND*.

An important question concerns the nature of these hp functions. There are different things that can be said about them. On the one hand, the hp function can be defined functionally. That is, we insist that to conform to hp-*COND* there must *be* some function (call it 'hp') such that for all times, t, $\text{cr}_t(\bullet) = \text{hp}(\bullet|E_t)$. On the other hand, one can see the hp function as something more substantive. Perhaps the hp function is some credence function that is the best representation of the agent's very early doxastic state. Or, if one is more of an objective Bayesian, then perhaps the hp function is the ideally rational credence function that one should have if one has no evidence. A different approach is to fix the hp function as some credence function that is the best representation of the agent's doxastic state at the beginning of the scenario that we wish

[7] Chris Meacham ([2007]) presents and briefly considers hp-*COND*.

to model. If we go this way, then it is not the *function* that is hypothetical, but rather its *prior*-ness that is hypothetical, since it is not *ultimately* prior.[8]

So, which approach should we go for? Different approaches have different virtues, but I think there are especially attractive features of the last approach. Recall, that we are attempting to evaluate the performance of human epistemic agents. This last approach allows us to do this in different ways depending on our purposes and keep irrelevant detail "off-stage". Say we want to model how a detective should respond to some new evidence learned during a morning briefing. We can model this by fixing his hp function as the best representation of his belief state as he came into the morning briefing. This allows us to ignore questions about whether or not his hp function is appropriate. Of course, if we want to ask questions about the appropriateness of this hp function, we can do so. But then we are modeling a new situation, and so will need to pick a different hp function, one that is the best representation of his belief state at some earlier time. This approach seems to give the variability that is appropriate to an evaluative epistemic theory.[9]

Note, however, a complication with this way of going. It is very plausible that there is nothing like a complete credence function that is the best representation of an agent's belief state at a time. Take the detective before he comes into the briefing. At best we will have enough information to only partially define a credence function at that time. But for hp-*COND* to put real constraints on how beliefs evolve, we must have a fairly robust hp function. To get this, I suggest that we adopt the functional approach

[8] See Meacham ([2007], pp. 5-7)for a brief summary of these sorts of proposals.
[9] A natural question for this approach is how we make sense of something like an "all things considered" epistemic 'ought'. Officially, I'd like to remain agnostic about this issue, and even over the existence of

constrained by the approach I favor. Thus, we take the partial constraints from looking at the agent's belief state and dispositions at the chosen prior time (t_0). We then say that there must be a function, hp, such that for all times, $t > t_0$, $cr_t(\bullet) = hp_{t0}(\bullet|E_t)$ where hp_{t0} meets the partial constraints garnered from the agent's belief state at t_0.

With hp-*COND* thus clarified, note some attractive things about understanding conditionalization in this way. First, hp-*COND* makes it clear that conditionalization is not offering us guidance in a way that sq-*COND* can seem to imply. Given our purposes here, that is a virtue. Second, and more importantly, hp-*COND* naturally allows for situations where an agent *loses* evidence, in a way that is impossible given sq-*COND*.[10]

5.5 Reliabilist Defeat

Adopting hp-*COND* opens up the possibility of modeling undercut evidence. But it does nothing to address the problem with Rigidity. To do this, we need to say something about how a proposition, *D*, can undercut a proposition, *H*, that does not require that $cr_0(H|D) = cr_0(H)$ and that $cr_1(H|D) < cr_1(H)$. It is here that RAE* has something to offer. As a first step, we need to say something about how we can look at the phenomenon of epistemic defeat from a reliabilist perspective.

Giving a completely general account of epistemic defeat is difficult. I will focus on giving a partial answer for the cases that concern the issues here. First, start with a partial reliabilist definition of what it is for a proposition to be a defeater of a belief:

---

such an epistemic 'ought'. However, a plausible way to model this would be to go all the way back to some very early doxastic state of the agent in question.

[10] hp-*COND* appears to be similar to Williamson's "ECOND", which he proposes in his ([2000]), p.220. However, it is not clear how Williamson understands the equivalent of the hp-function. Further, Williamson is concerned with what he calls 'evidential probability', which is *not* supposed to be the same thing as credence. Finally, Williamson seems to be attracted to ECOND for the possibility that it offers to respond to the problem of old evidence. This is not my motivation.

(Defeater) Proposition *D* is a defeater for the belief that *B*, relative to process *p* if: if *D* were true, then the process *p* that led to the belief that *B* would be unreliable.

Notice that a proposition is a defeater relative to a belief *and* a process that produced the belief, not relative to another proposition. So it makes no sense on this account to ask if one proposition is a defeater for another proposition. Note also that (Defeater) only provides a sufficient condition for being a defeater. It may be that there are other ways to be a defeater which do not depend on satisfying (Defeater). I leave this open.

(Defeater) suggests two different ways of evaluating whether or not a proposition is a defeater of a belief. The first way is to hold fixed the actual process *p* that led to the belief that *B*. Then we go to a world where *D* is true and see if the actual process *p* (now "transported" to a *D*-world) is unreliable. The second way is to go to the closest world where *D* is true and see if the process that leads to the belief that *B* in that closest *D*-world is a reliable process. In many situations, this process will be very similar (or even exactly the same) as the process *p*. But this will not be the case if *D* says something about the (actual) process *p*. Because it seems to correctly classify defeaters better, I opt for this second way of going.

To get the feel for (Defeater) some examples are helpful:

Example I:

B = "The wall is green."

Let the process that led to the belief that *B* be visual perception.

D = "The lighting is misleading."

D is thus a defeater for the belief that *B* since if the lighting were misleading, then this would render normal visual perception unreliable.

Example II:

B = "It is raining."

Let the process that led to the belief that *B* be visual perception.

D = "Your visual system is malfunctioning."

D is thus a defeater for the belief that *B* since if your visual system were malfunctioning, then this would render the process of visual perception (which led to the belief that *B*) unreliable.

Example III:

B = "There is a thunderstorm."

Let the process that led to the belief that *B* be auditory perception.

D = "You believe *B* because of visual perception and your auditory system is malfunctioning."

D is a defeater for the belief that *B* since if *D* were true, then you believed *B* because of a malfunctioning visual system. This process is unreliable.

Example IV:

B = "Ed said 'hello'."

Let the process that led to the belief that *B* be normal auditory perception.

D = "There is an Ed-impersonator in the office."

D is thus a defeater for the belief that *B* since if there were an Ed-impersonator in the office, this would render this kind of auditory perception unreliable.

Example V:

Leave *B* unspecified.

D = "The process that led to the belief that *B* is unreliable."

D is a defeater for the belief that *B* since if the process that led to the belief that *B* were unreliable, then (trivially) the process that led to the belief that *B* would be unreliable.

Note that an attractive feature of this (partial) definition of a defeater is that it allows a proposition to be a defeater of a belief even if the reason for the belief is non-doxastic. This is important given our motivation: we want to model situations where belief in an evidence proposition, which is evidence in virtue of some non-doxastic fact, is defeated.

Consider now the various kinds of belief-forming processes. When we are concerned with belief in evidence propositions, these will typically be perceptual processes. So, for instance, my belief that the light is on is formed by a process of human visual perception in a certain kind of environment. My belief that it is windy outside as I fall asleep is formed by a process of human auditory perception in a certain kind of environment. It is plausible that in certain situations, the relevant process of belief-formation involves other beliefs that the agent has. So, for instance, my belief that it is windy outside depends not purely on auditory perception, but also on beliefs

about what it sounds like when it is windy. Now, figuring out the correct process when assessing the reliability of a belief-forming process is a difficult thing to do. But the appropriate processes for such assessments will sometimes involve background beliefs that the agent in question holds. This need not always be the case, but it will sometimes be the case.

Consider, then, the following. Let *w* be the process of forming the belief that *B*. Let *D* be a defeater of *B* relative to *w*. Focus now on the process:

p: Forming the belief that *B* while strongly believing that *D*.[11]

I take it that processes like *p* are often the appropriate processes on which to focus when assessing reliability. Further, I maintain that this process is (for most people like us) an unreliable way of coming to believe *B*. To see why, consider two possibilities: either *D* is true or it is not.

(1) If *D* is true, then the process *p* that leads to the belief that *B* is almost certainly unreliable. If *D* is true, then *w* is unreliable (since this is what it is to be a defeater). Further, it is unlikely that adding the belief that *w* is unreliable somehow renders the total process, *p*, reliable.

(2) If *D* is false, then things are a bit more complicated. We want to know if the process *p* is reliable in such situations. We can usefully break this question into two parts. If *w* itself is unreliable, then *p* will be unreliable. Again, adding the belief that *D* to an unreliable process (*w*) surely does not give us a process that is reliable. However, this situation is not relevant to our concerns. For if *w* is

[11] I am assuming here that it makes sense to *add* something to a belief-forming process. One might worry about this. For instance, Schmitt ([1984]) does (p.8, footnote 16). But Schmitt is worried about adding some actual process to a merely possible process. In the text, all that I am doing is specifying the relevant

unreliable, then *B* is not evidence in the first place. So, we are focused on the scenario where *w* is reliable, and *D* is false. In this situation it seems that the total process *p* is reliable. That is, if we restrict our attention to those situations where *B* is evidence, and *D* is false, then *p* is reliable.

We want to know, however, whether *p* is a reliable process when we don't restrict things in this kind of way. To answer this, then, we need to know how often the agent is in situations in which he has false defeating beliefs. If the agent's beliefs in propositions like *D* are always false, then the reliability of process *p* is equivalent to the reliability of the more general process, *w*. But if the agent's beliefs in propositions like *D* are for the most part true, then the reliability of the process *p* will be less than the reliability of *w*. If that's all true of an agent, then adding a strong defeating belief to an otherwise reliable process yields an unreliable process (that is, shifting from *w* as the cause of the belief to *p* decreases the reliability).

Thus, if RAE* is adopted, RAE* provides one way in which an agent can lose evidence. Say that the belief that it is snowing is caused by a reliable perceptual process (*w*). Assuming it meets the other conditions in RAE*, that it is snowing is evidence. But if one then comes to have a strong belief in a defeater for this belief, the process leading to the belief that it is snowing is now a more specific process (*p*) with reduced reliability. Thus, that it is snowing is no longer evidence according to RAE*. Further, if (as we're assuming) this is the right way to type processes, the agent will not have available the process *w* that reliably indicates that it is snowing.

---

actual process more narrowly. But even if we were discussing merely possible processes, it still makes sense to specify such processes more or less narrowly, so I see nothing objectionable in this idea.

The following might help for expositional clarity: Let 'pr' be a probability function that is not to be understood as the agent's credence function, but is defined from the relevant frequency information. The reliabilist will think that there is something like such a probability function, since it will be needed for formulating facts about reliability (in fact, 'pr' made an appearance in Chapter 4 in the statement of RAE*). Given this, one can think of the proposal in the following way. When an agent is often accurate about propositions like *D*, or if the agent's beliefs in propositions like *D* are often false, but primarily when process *w* is itself unreliable, then the following will be true:

$$\text{pr}(B|w \wedge \text{strong belief that } D) = \text{pr}(B|p) < \text{pr}(B|w).$$

On the other hand, if an agent is very inaccurate about propositions like *D*, we'll have the following:

$$\text{pr}(B|w \wedge \text{strong belief that } D) = \text{pr}(B|p) \approx \text{pr}(\text{B}|w).$$

Thus, depending on general features of the agent in question, the addition of defeating beliefs to some process *w* can affect the reliability of the total belief-forming process.

Thus, we get the result that adding strong belief in a defeating proposition can impugn the reliability of an otherwise reliable belief-forming process. This, of course, is not *necessarily* true. Rather, it is contingent that defeating beliefs impugn reliability and thus can lead to evidence loss. If we consider agents that constantly form beliefs in defeating propositions like *D*, even though such situations do not obtain, then the presence of such strong beliefs is unlikely to affect the reliability of the belief-forming processes in question. But this seems to be the correct result, both intuitively, and according to reliabilist scruples. Such agents are what we might think of as overly-

skeptical, and it is inappropriate for overly-skeptical beliefs to impugn one's justification. Further, such skeptical beliefs do not affect the reliability of the belief-forming process, and so on reliabilist grounds, should be irrelevant to questions concerning the evidential status of a proposition.

5.5.1 Frederick Schmitt on Reliabilist Defeat

In his ([1984]), Frederick Schmitt takes up the issue of epistemic defeat from a reliabilist perspective. In particular, he is concerned to argue that in situations like the Red Jellybean Case, agents are no longer justified in believing the proposition which is delivered by a reliable process but for which there are skeptical background beliefs. Schmitt settles on the following proposal:

> "*S* is justified in believing that *P* only if there is no unexcercised *P*-withholding process *r′* whose reliability is at least as great as that of the process *r* that results in *S*'s belief that *P*." (p. 8, notation changed for continuity with this paper)

To put this kind of proposal in terms relevant to RAE*, the proposal would say that *P* is evidence for *S* just in case *P* is produced by a reliable belief-forming process, *r*, *and* there is no unexcercised *P*-withholding process *r′* whose reliability is at least as great as that of the process *r*.

Consider a case where I have skeptical background beliefs. Suppose that I am looking at a red apple in normal light, but strongly believe that the lighting is misleading. According to Schmitt, there is a reliable process leading to the belief that there is a red apple. This is the usual perceptual process. But there is also a reliable process leading me to withhold that belief. This is the process that begins with the belief

that the lighting is misleading and leads me to withhold beliefs about the colors of objects. In such a case, Schmitt says, the belief that the apple is red is not justified.

This could be carried over to RAE* as follows. An evidence proposition must not only be reliably formed (or formable), it also must not be such that there is another reliable process that leads one to withhold belief in that proposition. Thus, if we wanted to figure out an agent's evidence we would go through a three-step process:

1. We take the set of potential evidence propositions. This is supplied by something outside of RAE*, and could be simply the set of *all* propositions.
2. We then apply RAE* to rule out all those propositions that are not reliably formed (or formable).
3. We then further rule out all those propositions for which there are reliable belief-withholding processes.

My proposal is that we only look at the first two steps of this process, but realize that it is not such a simple thing to see which propositions are reliably formed. In particular, when looking at the red apple the usual perceptual process is reliable, but if I were to believe that the lighting is red, then the process I actually would employ would be: forming a belief with a defeating background belief. If I am appropriately sensitive to background conditions, then *this* process is not itself reliable. Accordingly, step two rules out from my evidence set the proposition that there is a red apple. Though true, it is not reliably formed.

I prefer my account to Schmitt's because it is not clear what a reliable *P*-withholding process is. I am currently withholding belief in *many* propositions. Am I

then executing all sorts of reliable *P*-withholding processes? I'm not sure. Since my proposal makes due without this notion, I think it is to be preferred.

5.6 A Reliabilist Solution to Undercut Evidence

We are finally in a position to see how RAE* can handle the phenomenon of undercut evidence. Say that (at t_1) Wally is looking with his well functioning eyes (connected to his well-functioning visual system) at the red jellybean under normal light. Grant that Wally is reliably connected to the truth of *R* = "The jellybean is red." According to RAE*, *R* is evidence. If one prefers RCE, this need not be the case, but assume that *R* meets the other conditions on having evidence to be part of Wally's evidence set. Accordingly:

$$cr_1(R) = hp(R|R) = 1.$$

But now imagine that Wally is told by an informant (at *t2*) that the lighting is abnormal: the lights are red, not white. Let *RL* be the proposition that asserts that the lights are red. Now, *RL* is not true. If it were, then it is possible that *R* would not be evidence since the process that led to the belief that *R* would be unreliable. Further, the informant may or may not be reliable. But *TRL* = "I am told *RL*," is true when considered by Wally. Given that Wally is a normal human, Wally is reliably connected with the truth of *TRL*. If one only adopts RCE, then we can assume that *TRL* meets the other conditions on being evidence. Thus, whether one endorses RAE* or RCE, it follows that *TRL* is part of Wally's evidence set, so:

$$cr_2(TRL) = hp(TRL|TRL) = 1.$$

Now there is a question: At *t2*, what is Wally's credence in *R*? RAE* and RCE are compatible with several different answers. If we assume that Wally believes that the

informant is reliable, then it is plausible that hp(*RL*|*TRL*) is quite high. By hp-*COND*, $cr_2(RL)$ is high. Thus, we have a situation where Wally has a strong belief in a proposition that is a defeater for *R*. Accordingly, the process that leads him to believe *R* at *t2* is not reliable. What is that process? It is the process of believing *R* based on visual perception with the addition of strong belief that the environment is misleading. Thus, it is the process of believing *R* while strongly believing a defeater for *R*. If this is an unreliable process, as I argued it will be for many normal humans, then *R* is not part of Wally's evidence set at *t2* since Wally does not have a reliable route to it. This is true whether one endorses RAE* or RCE. Accordingly, Wally's $cr_2(R) \neq 1$. Wally has lost evidence.

Note, however, that if Wally believes the informant to be unreliable, then we get a different verdict. If he believes the informant to be unreliable, then it is plausible that hp(*RL*|*TRL*) is quite low. Accordingly, $cr_2(RL)$ is low. Thus, this is not a situation where Wally has a strong belief in a defeating proposition. Wally *does* have a strong belief in *TRL*, but given the account of defeat given above, this is not a defeater for his belief that *R*. Accordingly, *R* is still in Wally's evidence set at *t2*. Opinions may be split over this result, but it seems to me to be an attractive one. For we don't think that Wally loses evidence if Pete, the office's practical jokester tells Wally that the lighting is misleading. But the situation is different if Abe, the honest and stoic accountant, tells him this.

Consider now a variant of the original case. This time, Wally is first told that there are red lights, and then looks at the jellybean. Both RAE* and RCE give a plausible result in this case, too. At *t1*, Wally is told by the informant that the lights are

red. *TRL* is thus evidence for him, and if we assume that Abe is the informant, $cr_2(RL)$ is high since hp(*RL*|*TRL*) is high. Then, at *t2*, Wally looks at the red jellybean. For the same reason as above, Wally does not have a reliable route to the truth of *R* since he is harboring a strong defeating belief (*RL*). Thus, *R* is not evidence for Wally.

We can represent both cases where Wally believes the informant to be reliable in the following tables. I have included the proposition *AR* = "It appears as if there is a red jellybean," to give more detail about these cases:

Table 2: Original Case – Undercutting Belief

Time	*R*	*AR*	*TRL*	*RL*	EVIDENCE
t0	cr(*R*) = *n* = low	cr(*AR*) = low	cr(*TRL*) = low	cr(*RL*) = low	{∅}
t1	cr(*R*) = 1	cr(*AR*) = 1	cr(*TRL*) = low	cr(*RL*) = low	{*R, AR*}
t2	cr(*R*) = hp(*R*|*TRL, AR*) > *n* (but less than 1)	cr(*AR*) = 1	cr(*TRL*) = 1	cr(*RL*) = high	{*AR, TRL*}

Table 3: Reversed Case – Skeptical Background Belief

Time	*R*	*AR*	*TRL*	*RL*	EVIDENCE
t0	cr(*R*) = *n* = low	cr(*AR*) = low	cr(*TRL*) = low	cr(*RL*) = low	{∅}
t1	cr(*R*) = *n* = low	cr(*AR*) = low	cr(*TRL*) = 1	cr(*RL*) = high	{*TRL*}
t2	cr(*R*) = hp(*R*|*TRL, AR*) > *n* (but less than 1)	cr(*AR*) = 1	cr(*TRL*) = 1	cr(*RL*) = high	{*AR, TRL*}

Note that in these tables $cr_2(R)$ does not return to its value at *t1*. This is because the red appearance, even after being told that there are red lights is likely to increase the probability that the jellybean is red. Other than this complication, the tables represent the stories as I told them above.

We see, then, that RAE* or RCE together with hp-*COND* and a reliabilist account of defeat, allows one to handle cases of undercut evidence. In general, here is how the solution works: hp-*COND* permits evidence to be lost. The reliabilist account of defeat allows that a proposition can be an undercutting defeater of a *non-doxastic*

reason. This is because the process of believing a proposition while strongly believing a defeating proposition is unreliable. RAE*/RCE say that only reliable processes yield evidence.

It is important to point out that one could embrace a solution to the problem of undercut evidence that is not committed to reliabilism about evidence in general or RAE*/RCE in particular. The key features of a solution are:

(1) the adoption of hp-*COND*,

(2) the adoption of some account of defeat that allows a proposition's non-doxastic reasons to be defeated, and

(3) the adoption of an account of evidence that does not allow a defeated proposition to be evidence.

None of (1) – (3) make essential reference to reliability, so a solution could be offered that does not adopt RAE*/RCE. However, it is quite natural to appeal to reliability to handle (2), since it is typically something about the process that led to the belief that is being defeated in situations of non-doxastic defeat. Further, simply stipulating that defeated propositions are not evidence is a theoretically unattractive option. RAE*/RCE gives an epistemically relevant reason—reliability—why defeated propositions might not be evidence.

In closing this section, I note two features about this model of undercut evidence that I've left unspecified. First, I've left unspecified how we are to select the relevant processes of belief-formation over other processes. Answering this, of course, runs us head-on into the generality problem. In Chapter 3, Section 5 I discuss my preferred responses to the generality problem, but my model in this chapter is compatible with

many different responses. A second part of the model that I've left unspecified concerns the precise meaning of the term 'strong belief'. A natural specification of the proposal is to take a strongly believed proposition to be any proposition with credence greater than 0.5. Thus, if you think that it is more likely than not that a defeating proposition is true, then this is sufficient to impugn the reliability of the process that leads to the belief in question. Though this may be a good general guide, I hesitate to associate a particular credence value with the term 'strong belief'. This is because what we care about is the overall reliability of the process in question. It may be that for an agent there is no significant drop in reliability until credence in a defeating proposition is greater than 0.6. In such a case, associating strong belief with credence greater than 0.5 will get an incorrect result. Thus, I propose that we let 'strong belief' vary so as to track the reliability of the process in question.[12]

5.7 A Different Way to Lose Evidence

It will help to clarify the model I've presented for undercut evidence to note the general strategy being employed. The general claim is that reliability considerations alone are sufficient to explain how and when evidence is lost. In the previous sections, I've attempted to show how this works for a certain class of situations: the class of situations where an agent comes to strongly believe an undercutting defeater (in the sense defined) for a proposition in the agent's evidence set. Since many have strong intuitions that in such a situation one *can* lose evidence, I have attempted to show how

[12] As a side note, it is important to realize that a process need not become completely unreliable for it to fail to yield evidence propositions. Perhaps, as seems plausible, the standard of reliability for evidence propositions imposed by RAE*/RCE is quite high. Then, a small decrease in the reliability of the process will be sufficient to expunge a proposition from the evidence set. This can be so, even if the process that leads to the now-removed evidence proposition is still quite reliable.

reliability considerations can account for this. What I've shown is that in many situations, adding a strong belief in a defeater will lower the reliability of the process leading to the belief in the evidence proposition. Given RAE* or RCE, this is sufficient to knock that evidence proposition out of the evidence set.

Given this way of looking at the general strategy, it is useful to consider a different class of situations where it is natural to think that an agent can lose evidence. This class of situations is similar, though distinct from the ones we have considered so far. The situations considered so far are situations where an agent comes to strongly believe a defeater for a proposition in the agent's evidence set. The class of situations I would now like to consider are those situations where an agent gets conflicting reports from two very reliable sources.

Consider an example. I wake up in the morning and read in the *New York Times* that Obama won the election. If we're not imposing too high a reliability constraint on evidence, then it seems plausible that "Obama won the election" (*O*) is a member of my evidence set. But now imagine that on the subway I happen to see a copy of the *Washington Post* which reports that Obama didn't win the election. Grant that the reliability of the New York Times and the Washington Post are high and roughly equivalent. It seems very plausible that in such a situation *O* should no longer be in my evidence set.

This is a case, however, that the account sketched above does not address. It's not the case that $\neg O$ could be in my evidence set upon looking at the *Post* both because this would render me incoherent, and because to allow this in my evidence set would be to make a similar mistake to allowing *O* to be in my evidence set. One might point out

that nevertheless, the proposition $wp\neg O$: "the Washington Post says Obama didn't win the election," is plausibly part of my evidence set. One might think that having $wp\neg O$ in my evidence set will knock O out. This, however, is not the case. The account of defeated evidence that I gave above doesn't have this effect. The proposition $wp\neg O$ is not a defeater of my belief in O in the sense that I defined above. To be a defeater, the truth of $wp\neg O$ would have to render the process leading to the belief that O unreliable. But that's not the case here. The *Times* could be very reliable even if it is true that the equally reliable *Post* reports the opposite.

Notice that this kind of situation is bound to arise often. The general structure of the case is where a reliable source indicates P and a different reliable source indicates $\neg P$. The sources could be two newspapers, or vision and memory, or testimony and memory, or audition and vision, etc. Further, note that this is an especially important issue for a defender of RAE* over RAE*-t. For, at first glance, RAE* would seem to say that in such a situation both P and $\neg P$ are evidence. But this would result in $cr(P) = cr(\neg P) = 1$, which is incoherent.

Though we can't handle these cases in the way that we handled undercutting defeat above, I think there is an even simpler explanation—appealing only to reliability considerations—for why in such a situation neither P nor $\neg P$ are in the agent's evidence set. In such a situation, the process that leads to either the belief that P or $\neg P$ is unreliable. It is the process of believing in the face of reliable conflicting testimony, and unless one has bizarre faculties, this will be unreliable. Informally, if there are conflicting reports, then one can choose to believe one or the other. But given no extra

information, one is just as likely to choose correctly as incorrectly. So the process is not a reliable one.

To see this more precisely, consider the case of the newspapers, again. For simplicity let's grant that in such a situation the two newspapers are equally reliable, in the sense that for any proposition X:

$$\text{pr}(nyX|X) = \text{pr}(wpX|X) = p$$

and

$$\text{pr}(nyX|\neg X) = \text{pr}(wpX|\neg X) = q$$

where p is high and q is low. Further, for simplicity, grant that $q = 1 - p$.[13] Thus, the *New York Times* and the *Washington Post* are both reliable sources, in the sense that:

$$\text{pr}(X|nyX) = \frac{\text{pr}(X) \times \text{pr}(\text{ny}X|X)}{[\text{pr}(X) \times \text{pr}(\text{ny}X|X)] + [\text{pr}(\neg X) \times \text{pr}(\text{ny}X|\neg X)]} = \frac{\text{pr}(X)}{\text{pr}(X) + [(q/p) \times \text{pr}(\neg X)]}$$

Notice that as p increases, q decreases, $\text{pr}(X|nyX)$ approaches 1. So, if p is high for the *Times* and for the *Post*, then $\text{pr}(X|nyX) = \text{pr}(X|wpX)$ = high (depending on the prior probability of the thing reported.[14] But look what happens if the *Times* and the *Post* give contradictory reports. We're interested in:

$$\text{pr}(X|nyX \wedge wp\neg X) = \frac{\text{pr}(X) \times \text{pr}(\text{ny}X \wedge \text{wp}\neg X|X)}{[\text{pr}(X) \times \text{pr}(\text{ny}X \wedge \text{wp}\neg X|X)] + [\text{pr}(\neg X) \times \text{pr}(\text{ny}X \wedge \text{wp}\neg X|\neg X)]}$$

If $\text{pr}(nyX|X)$ and $\text{pr}(wp\neg X|X)$ are independent of each other, then $\text{pr}(nyX|wp\neg X \wedge X) = \text{pr}(nyX|X)$. Although somewhat of an idealization, this is plausible. It says that, assuming X is true, simply learning that the *Post* reported $\neg X$ doesn't alter the

[13] This amounts to the assumption that $\text{pr}(nyX|\neg X) = \text{pr}(ny\neg X|X)$ and $\text{pr}(wpP|\neg P) = \text{pr}(wp\neg P|P)$.
[14] In order for p to be high, we might need to restrict our attention to "newsworthy propositions". For presumably, p is quite low if we include all propositions, since most true propositions aren't reported in newspapers.

probability that the *Times* reports *X*. Assuming that the *Times* and the *Post* don't share reporters, copy each other's articles, or share a source, this is a reasonable assumption. Granting this it follows that:

$$\text{pr}(X|nyX \wedge \text{wp}\neg X) = \frac{\text{pr}(X) \times pq}{[\text{pr}(X) \times pq] + [\text{pr}(\neg X) \times pq]} = \text{pr}(X)$$

So, conditional on receiving contradictory reports from two reliable newspapers, the probability of *X* is just the prior probability of *X*. So, if before seeing either report *X* wasn't very probable, then after getting two contradictory reports it isn't very probable. This is what underlies the idea that if I form a belief that *X* on the basis of the *Times* reporting *X* and the *Post* reporting ¬*X*, I'll not be using a reliable process of belief formation.

This brief detour shows that the general strategy employed here can be deployed to handle cases of lost evidence that do not fall under the rather specific class of situations where some evidence proposition suffers non-doxastic defeat.

5.8 Trouble for the Solution

The last section clarified the general strategy, and noted how it can be deployed to handle different situations. It is now time to turn our attention back to non-doxastic defeat. Though the account I've given for handling such situations works well in the Red Jellybean Case and others like it, it faces a surprising difficulty in the snow case that began the chapter. This is interesting since the snow case seems to be structurally similar to the Red Jellybean Case. To bring out the difficulty, I will first recount the snow case with a bit more detail.

The Snow Case:

At *t1* Wally is looking with his well-functioning eyes (connected to his well-functioning visual system) at the snow falling outside the window. Grant that Wally is reliably connected to the truth of *S* = "It is snowing." Thus, according to RAE* and RAE*-t this is a member of Wally's evidence set at *t1*. As long as *S* meets the other conditions on being evidence, an adherent of RCE would grant this, too. Accordingly: $cr_1(S) = hp(S|S) = 1$.

But now imagine that Wally is told by an informant (at *t2*) that someone is playing a prank on him, dropping fake snow from the window above so that it looks as if it is snowing. Let *F* refer to the proposition that asserts that someone is doing this. Now, *F* may or may not be true, and the informant may or may not be reliable. But *TF* = "I am told *F*," is true when considered by Wally. Given that Wally is a normal human, Wally is reliably connected with the truth of *TF*. Again, given the usual stipulations, it follows that *TF* is part of Wally's evidence set, so: $cr_2(TF) = hp(TF|TF) = 1$.

This is the Snow Case. Assuming that Wally believes the informant to be reliable, we now face a question: At *t2*, what is Wally's credence in *S*? The defeating proposition in this case is *F*. This case seems to be structurally similar to the Red Jellybean Case above, and so it is natural to think that $cr_2(F)$ is high for just the same reason that $cr_2(RL)$ was high. But there is an important difference between the cases. In particular, it is not so clear that $cr_2(F)$ *is* high in this situation. Why is this unclear? Because although it is plausible that $hp(F|TF)$ = high (since Wally believes the informant to be reliable), it is also plausible that $hp(F|TF \wedge S)$ = low. That is, conditional on it snowing and being told that it is a prank, it is not likely that there is a

prank. After all, who tries to fake a snowstorm *during a snowstorm*? But if that's right then something odd is going on. It seems that it is consistent with the view being sketched that *S* is still part of Wally's evidence set at *t2*, even in light of the new information. For if *S* is evidence, then $cr_2(S) = 1$, and so $cr_2(F)$ is not high. And if that's the case, then there is still a reliable route to the truth of *S*. Of course, it could also go the other way. *S* could drop out of the evidence set at *t2*, in which case Wally strongly believes *F* and so there is no reliable route to the truth of *S*.

It seems, then, that both the following are consistent with the story and the account of undercut evidence that I have offered:

Table 4: Snow Case I

Time	*S*	*AS*	*TF*	*F*	Evidence
t0	cr(*S*) = *n*	cr(*AS*) = low	cr(*TF*) = low	cr(*F*) = low	{∅}
t1	cr(*S*) = 1	cr(*AS*) = 1	cr(*TF*) = low	cr(*F*) = low	{*S, AS*}
t2	cr(*S*) = hp(*S*|*TF, AS*) > *n*	cr(*AS*) = 1	cr(*TF*) = 1	cr(*F*) = high	{*AS, TF*}

Table 5: Snow Case II

Time	*S*	*AS*	*TF*	*F*	Evidence
t0	cr(*S*) = *n* = low	cr(*AS*) = low	cr(*TF*) = low	cr(*F*) = low	{∅}
t1	cr(*S*) = 1	cr(*AS*) = 1	cr(*TF*) = low	cr(*F*) = low	{*S, AS*}
t2	cr(*S*) = 1	cr(*AS*) = 1	cr(*TF*) = 1	cr(*F*) = hp(*F*|*S, AS, TF*) = low	{*S, AS, TF*}

So, just what is causing the difference between the Red Jellybean Case and the Snow Case? There is one significant difference. In the Red Jellybean Case, it is plausible that

hp(*RL*|*TRL*) = high

but it is also plausible that

hp(*RL*|*TRL* ∧ *R*) = high.

Accordingly, credence in *RL* is high whether or not we think that *R* is in the evidence set or not. But in the Snow Case, the situation is different. Whether or not *S* is in the

evidence set seems to affect whether or not the defeating proposition is strongly believed. The general phenomenon here is that both the following claims are true, and yet in tension with each other:

(1) One's beliefs at a time depend in part on one's evidence at that time.

(2) One's evidence at a time depends in part on one's beliefs at that time.

From (1), whether or not *F* is strongly believed at *t2* depends on whether or not *S* is evidence at *t2*. From (2), whether or not *S* is evidence at *t2* depends on whether or not *F* is strongly believed. The Red Jellybean Case sidesteps this difficulty, because *RL* should be strongly believed (according to the story) whether or not *R* is in the evidence set. But in the Snow Case, things are not so simple.

What we see, then, is that in certain cases (e.g., the Red Jellybean Case) the proposal I've given issues a clear verdict about the epistemically appropriate response. However, in certain other cases (e.g., the Snow Case), the proposal I've given does not issue one clear verdict about the epistemically appropriate response.

It is worth inquiring whether this is a special oddity that my account faces, or if other accounts would face this too. If my diagnosis of the source of the oddity is correct, and if an account allows defeated and lost evidence, then this kind of phenomenon will be common. The root of the oddity is the interdependence between beliefs and evidence. Note that in most Bayesian stories, one's evidence is simply *stipulated*. Of course, stipulational "accounts" of evidence will not face this oddity, but that is merely because the problem is being ignored, not because it is being confronted and solved.

5.9 The Holistic Constraint on Evidence

One response to this situation is to simply leave things indeterminate. According to this response, in cases like the Snow Case, there is more than one epistemically appropriate response for the agent. For in such cases we have a situation where what your evidence is depends on your beliefs, and what your beliefs are depend on your evidence. There is an air of Quinean holism[15] to this situation, and perhaps we shouldn't expect *one* unequivocal verdict in such cases.

I'm uncertain as to the acceptability of such a response. In light of this, I will propose a further reliabilist constraint that attempts to determine which response is the correct response in indeterminate cases like the Snow Case. The proposed constraint resolves the indeterminacy when there is indeterminacy. The proposed constraint is given by the following test:

The Holistic Constraint on Evidence (HCE)

1. Take all the propositions $P_1, P_2, \ldots, P_n$ in the candidate evidence set at t (call this set $\{P_i\}$).

2. Remove one proposition, P_j from $\{P_i\}$.

3. Conditionalize the hp-function on $\{P_i\} - P_j$.

4. Are there any propositions, X such that $\text{hp}(X|\ \{P_i\} - P_j) > n$ and where X is a defeater of one of the P_i?

5. If so, then the account that gives $\{P_i\}$ as one's evidence set is incorrect.

6. If not, then go to step 2 and repeat with some other member of $\{P_i\}$.

[15] See, for instance, Quine & Ullian ([1978]).

7. If the answer at step 4 is 'no' for all members of $\{P_i\}$, then $\{P_i\}$ passes the holistic constraint on evidence.

One can then say that all evidence sets must pass HCE.

There are several things about HCE that must be explained. First, this test assumes the we can partition evidence into propositional components. This is a sizable assumption, but it is one that is necessary for any kind of reliabilist account of evidence. This might give one reason to dislike reliability conditions in general, but it is not a *specific* objection to HCE.

Second, and related to this, one might worry that the verdicts that the HCE test yields will be sensitive to features of how we individuate propositions. For instance, imagine a situation where it is most natural to describe the candidate evidence set as containing three propositions: $\{P1, P2, P3\}$. Further, it is true that if *P1* is removed, then there is a defeater for *P2*, however this is not the case if *P1* and *P3* are removed. If that's the case, then it seems we could just re-describe the candidate evidence set as containing *two* propositions: $\{(P1 \wedge P3), P2\}$. Although the first candidate evidence set failed the HCE test, this second (equivalent) evidence set passes the test. This unhappy consequence can be avoided, however, by noting that this is only a necessary condition on evidence sets, so there may be other reasons to rule out one or the other of these evidence sets. In particular, RAE* seems well-placed to individuate evidence propositions here. RAE*, recall, allows propositions into the evidence set when there is a reliable process that yields belief in that proposition. We might then say that *P1* and *P3* are to be treated as separate propositions when there are distinct reliable processes

yielding belief in each of them. If there are not distinct reliable processes, but rather one, then (*P1* ∧ *P3*) is to be treated as one proposition.

A final note about HCE concerns step 2. As I state it, we take one proposition out at a time, and then check to see if there are any defeaters. But one might think that it is also necessary to take out more than one proposition at a time and check to see if there are any defeaters of the potential evidence set. That is, suppose again that {*P1*, *P2*, *P3*} is the candidate evidence set. As I have it now, we take out *P1*, and then conditionalize on {*P2*, *P3*} to see if there are any defeaters. Then we take out *P2*, and conditionalize on {*P1* ∧ *P3*} to see if there are any defeaters, etc. But as I've stated step 2, we do not take out *P1* and *P2* and then conditionalize on {*P3*} to see if there are any defeaters. It seems that there could be no reason to take out one proposition at a time, but not multiple propositions. Nevertheless, I have not come across cases that demand this, so while I would not be against modifying step 2 in this way, I do not see any need to do so at present.

Now let's see how HCE helps to solve the problem that arose in the Snow Case. At *t2* we seemed to have two candidate evidence sets. HCE, however, says that Wally's evidence set at *t2* cannot be {*S, AS, TF*}. This is because at *t2*, the evidence set fails the HCE test. If we remove *S* from {*S, AS, TF*}, we are left with {*AS, TF*}. If we then conditionalize the hp-function on {*AS, TF*}, we get hp(*F*|*AS,TF*) = high, where *F* is a defeater of *S*. Thus, the test is failed. The evidence set {*AS, TF*}, however, passes the HCE test, and so is acceptable. Note that if *TF* is taken out of the evidence set, then this might yield a strong belief in *S* since it is plausible that hp(*S*|*AS*) = high. However, *S* is not a defeater of *TF*, and so we have no violation of HCE.

If we apply the principles I've specified in one way (Snow Case I), then Wally's evidence set at *t2* is {*AS, TF*}; if we apply the principles differently (Snow Case II), then Wally's evidence set at *t2* is {*S, AS, TF*}. HCE decides the issue in favor of Snow Case I. So, application of HCE privileges one way of applying the principles I've articulated over the other. Further, it seems to me that it *is* correct to privilege Snow Case I over Snow Case II, given the scenario. Consider the situation: Wally is told by an informant that he regards as reliable that the snowstorm is being faked. It seems appropriate in such a situation for Wally to decrease his confidence that it is snowing. Accordingly, *S* should not be in Wally's evidence set. This is the result that HCE delivers.

5.10 Possible Counterexamples to HCE

One might worry about a certain feature of the HCE test. In particular, it assesses acceptable evidence sets in a piecemeal fashion. This feature might be thought to lead to counterintuitive results. In particular, one might be worried because there can be situations where any subset of one's total evidence set supports radically different propositions than one's entire evidence set. For example, consider the following example (from Kotzen [*ms*]). I have as evidence both that a prospective student specializes in metaphysics (*M*) and in epistemology (*E*). However, assume that there are three universities, UM, UE, and UME. UM specializes in metaphysics, UE specializes in epistemology, and UME specializes in both. Accordingly, cr(*UM*|*M*) = high, cr(*UE*|*E*) = high, cr(*UME*|*M* ∧ *E*) = high, cr(*UM*|*M* ∧ *E*) = cr(*UE*|*M* ∧ *E*) = low. In this situation, any subset of my evidence {*M*, *E*} will give a high credence to either *UM* or

UE, whereas the total evidence {*M*, *E*} will give both *UM* and *UE* low credence. This is, of course, a general phenomenon.

One might worry that we can leverage this into a counterexample against HCE. To be effective, the counterexample would need to give an evidence set and an associated hp function that seems intuitively acceptable. It must in addition have the feature that if we remove one of the propositions from the evidence set, we get strong belief in a defeater for some proposition in the evidence set. The structure of such a counterexample is as follows:

1. hp(*B*|*EB*) = high.

2. *B* defeats *EA*.

3. hp(*B*|*EA* ∧ *EB*) = low.

1 and 2 tell us that if you remove *EA* from the evidence set {*EA*, *EB*}, the result is a strong belief in a defeater for *EA*. HCE would thus say that {*EA*, *EB*} cannot be the agent's evidence set. 3 tells us that the defeater for *EA* is not strongly believed if the evidence set is {*EA*, *EB*}. This is necessary because if the defeater for *EA* is strongly believed when {*EA*, *EB*} is the evidence set, then RAE* will say that {*EA*, *EB*} isn't even a possible evidence set. So, if 1 – 3 are satisfied and if {*EA*, *EB*} is an intuitively correct evidence set, then there would be a counterexample to HCE.

One such case where 1 – 3 are true is the Snow Case, where *B* = *F*, *EB* = *TF*, and *EA* = *S*. However, in this case, the evidence set is not intuitively acceptable. We can render it intuitively acceptable by assuming that *S* is in the evidence set in some other way that does not depend on observations that would be unreliable if *F* were true. Then

we seem to have an acceptable evidence set, but we fail condition 2. *F* is no longer a defeater for *S*.

One might think that a slight variant of the Snow Case can cause trouble, however. Suppose that I come to the office with a strong belief that there will be people faking a snowstorm on the rooftop (*F*). However, in my office I've rigged up a video camera that is sending me a live feed from the roof. I see that the video camera shows no unusual activity. Grant that this proposition, that the video camera shows no unusual activity (*NUA*), is evidence. It is plausible that hp(*F*|*NUA*) = low. Now, suppose that after getting *NUA* as evidence, I see it snowing outside my window. It seems that *S* is evidence for me, so that both *NUA* and *S* are in my evidence set. One might think that HCE rules this out. For if you remove *NUA* from my evidence set, then it seems as though *F*, which is a defeater for *S*, is strongly believed. But this is unintuitive, since it seems as though there is nothing wrong with *S* and *NUA* being my evidence. Appearances to the contrary, however, this is not a problem for HCE. The candidate evidence set is {*S*, *NUA*}. HCE rules out this evidence set if, upon removing *NUA,* there is a strongly believed defeater for either *S* or *NUA*. But note that there isn't. For upon removing *NUA*, we must see what the value of *F* is conditional on the other evidence. That is, we must see what the value of hp(*F*|*S*) is. But, it is reasonable to think that hp(*F*|*S*) is low (since, again, who fakes a snowstorm *in a snowstorm?*). So, in the scenario considered, HCE allows an evidence set that contains *S* and *NUA*.

One might try to construct a more complicated kind of counterexample to cause problems for HCE. Consider one built off the university example just given. Here's the relevant structure:

4. hp($A|EA$) = hp($B|EB$) = high.

5. A defeats EB, B defeats EA.

6. hp($A|EA \wedge EB$) = hp($B|EA \wedge EB$) = low.

Let A = A committed the crime, EA = A is identified by the witness, B = B committed the crime, and EB = B is identified by the witness. Let the agent in question be a detective. We can tell the story in a way in which it is plausible that both EA and EB are the detective's evidence. Perhaps he heard the witness identify both A and B. In this case it is plausible that 4 and 6 are satisfied. But in this case 5 is false. A defeats EB just in case if A were true, then the process that led to the belief in EB is unreliable. So, if it were true that A committed the crime, is the process that led the detective to believe that B is identified by the witness unreliable? No. The process that led to that was something like regular perceptual processes, which aren't impugned if we change who committed the crime. So, this is not a counterexample either.

As one last attempt, we can try to rig things up so that 5 *is* satisfied. Let A = the EB-forming process is unreliable, EA = I am told A, B = the EA-forming process is unreliable, and EB = I am told B. It is plausible that hp($A|EA$) = hp($B|EB$) = high. We can tell the story in such a way that the informant both is reliable and believed to be. This makes 4 true. Further, by construction, A defeats EB, and B defeats EA. So 5 is true. Further, it seems that it is possible that EA and EB are both evidence for an agent. After all, EA and EB simply record that the agent was told something.

So, we have a counterexample to HCE if 6 is true. This would be true if hp($A|EA \wedge EB$) = hp($B|EA \wedge EB$) = low. This is a difficult case to think about, but I think it is plausible that 6 is not true. Consider just hp($B|EA \wedge EB$). This is the

confidence given to *B*, conditional on being told that *B* (*EB*) and being told that *A* (*EA*). If 6 is true, then the agent must assign a high degree of confidence to ¬*B* conditional on this. But *B* says that the process leading to the belief that *EA* is unreliable, so ¬*B* says that the process leading to the belief that *EA* is reliable. But this tells directly against what the agent was told and what is also part of his evidence (*EB*). So, I find it implausible that 6 is satisfied in this case. Thus, this is no counterexample.

I conclude that there are no obvious counterexamples to HCE. While far from a proof that there will be no such counterexamples, the unsuccessful search is suggestive.

5.11 Conclusion

In this chapter I've both clarified what a reliabilist account of evidence might look like, and showed how it is capable of doing some real work. The clarification of the account of evidence was to adopt hp-*COND* rather than sq-*COND*, which shows how evidence once acquired need not be kept forever. However, this chapter did more than this. By combining RAE* (or RCE) with hp-*COND* and a reliabilist account of defeat, we are able to make sense of situations in which one loses evidence in a particular way: because it is defeated by other beliefs that one holds. The account that I defended then faced a possible problem in that it allowed agents to retain evidence that it may seem they should not. This was brought out in the Snow Case. However, I proposed HCE, an extra constraint on evidence, which gives a solution to this problem. What we have, then, is an intuitively satisfying Bayesian account of how evidence can be undercut.

CHAPTER 6

DUTCH BOOKS AND CERTAIN LOSS

6.1 Introduction

Dutch Book Arguments (DBAs) have long been given for various Bayesian norms.[1] These include plausible norms like:

***PROB*:** The function cr(•) is rational only if cr(•) is a probability function.

***COND*:** $cr_{t1}(\bullet) = cr_{t0}(\bullet|E_{t1})$, where E_{t1} is the evidence at *t1*.

But they have also been given for less plausible[2] norms like:

***Reflection*[3]:** $cr_t(A|\langle cr_{t+}(A) = n\rangle) = n$ (for $t+$ later than t)

***Self-Respect*[4]:** $cr_t(A|\langle cr_t(A) = n\rangle) = n$

This raises something of a puzzle for those who want to use DBAs to argue for the plausible norms, but not the implausible ones.

Recently, Rachel Briggs ([2009]) has proposed a way to distinguish the DBAs for *PROB* and *COND* as good and the DBA for *Reflection* as bad. Her proposal bears close resemblance to a proposal given by David Christensen ([2007]), which aims to show that the DBA for *Self-Respect* is flawed. Christensen, however, holds that this flaw with the *Self-Respect* DBA infects the DBAs for *COND* and *Reflection*, but not the

[1] They go back at least to F. P. Ramsey's classic paper, "Truth and Probability" ([1926/1990]). The argument can be found other places, including de Finetti ([1937/1980]), Teller ([1973]), Skyrms ([1975]), Horwich ([1982]), Skyrms ([1987]), Sobel ([1990]), Armendt ([1992]), Skyrms ([1993]), Howson & Urbach ([1993]), Lewis ([1999]), McGee ([1999]) and Briggs ([2009]).
[2] For rejections of *Reflection* see, for instance, Christensen ([1991]), Arntzenius, ([2003]), Briggs ([2009]). For rejections of *Self-Respect* see, for instance, Christensen ([2007]), Roush, ([2009]).
[3] For details about *Reflection*, see van Fraassen ([1984], [1995]).
[4] This is simply the synchronic version of *Reflection*. The name '*Self-Respect*' comes from Christensen ([2007]).

one for *PROB*.[5] In this chapter, I'll be primarily concerned with the DBAs for *COND*, *Reflection*, and *Self-Respect*, leaving the DBAs for *PROB* aside.[6] Thus, Briggs maintains:

Good DBAs	Bad DBAs
COND	*Reflection*, *Self-Respect*

While Christensen claims:

Good DBAs	Bad DBAs
∅	*COND*, *Reflection*, *Self-Respect*

In this chapter I will argue that Briggs's classification can be defended, but not in the way that she proposes. To reach this conclusion, I'll first get clear on what the Christensen/Briggs response to these DBAs is. Next I'll show that Briggs's claim that this allows us to accept the DBA for *COND* and reject the DBAs for *Reflection* and *Self-Respect* is mistaken. However, I will then propose a modified way of distinguishing DBAs that builds off the Christensen/Briggs response. I will argue that this modified approach is plausible and that it succeeds in getting us the classification that Briggs desires. This shows what we would have hoped was true all along: there is a way to understand DBAs such that the plausible principles receive support from them, but not the implausible ones. However, in pursuing such an account, it will become especially clear the large role that certain assumptions play in Dutch Book Arguments. The proposal that I will give says nothing about how to understand the role that these assumptions play in DBAs, although it does make their importance more obvious. I will

[5] See Christensen ([2007]), footnote 3. The full story about Christensen's views on the matter is somewhat more complicated. In his ([1991]), he appears to see all diachronic Dutch Books as flawed, which puts the *COND* DBA and *Reflection* DBA in the same boat. In his ([1996]), however, Christensen suggests a way in which they might be distinguished. See Vineberg ([1997]) for criticism.

thus conclude this chapter by discussing this issue, and what bearing it has on Dutch Book Arguments.

Before beginning, it is perhaps worth noting something I won't be doing: I won't be giving much in the way of argument for the claim that *Reflection* and *Self-Respect* aren't norms of rationality. If one held the view that they were, then presumably one would not necessarily find it a welcome result that the DBAs for these principles come out as bad, while the DBA for *COND* comes out as good. Further, I won't be giving a general defense of *COND* as the only rational update rule. My project here is more modest. It is to show that there is a plausible way of understanding Dutch Book Arguments and according to this way of understanding them, there is a flaw in the DBAs for *Reflection* and *Self-Respect*, but not in the DBA for *COND*. Before we can get to this, however, some preliminary comments about DBAs are necessary.[7]

6.2 Dutch Book Arguments

I'll follow Hájek ([2008]) in distinguishing between a *Dutch Book*, a *Dutch Book Theorem*, and a *Dutch Book Argument*. A Dutch Book is simply a set of bets such that accepting those bets guarantees the bettor a loss. The Dutch Book Theorem is a statement of the following form:

> If an agent violates principle *p*, then there exists a set of bets such that the agent accepts all the bets and the bets make up a Dutch Book.

---

[6] Thus, when considering the DBAs, I'll be only considering agents who satisfy *PROB*.

[7] A terminological note: some writers distinguish between *Dutch Book Arguments* and *Dutch Strategy Arguments*. Roughly, a Dutch Book is a sequence of bets offered at one time, that result in a certain loss or gain for the agent. A Dutch Strategy is a sequence of bets offered at different times, that together result in a certain loss or gain for the agent. This terminology is rendered ambiguous, however, because of the argument for *Self-Respect* : the bets are offered all at one time, but the way in which the bets are set up

Using the Dutch Book Theorem we then construct the Dutch Book Argument. This argument says that you shouldn't violate principle *p*, since doing so will leave you vulnerable to certain loss of money.

Before we can give such an argument, however, we'll need some connection between the acceptance of bets and credences. The standard connection is as follows: if cr(*P*) = *n*, an agent should be willing to pay $n \times \$D$ for a bet on *P* that pays $\$D$ if *P* is true. If I am such that cr(*Heads*) = 0.2, I should be willing to pay $2 for a bet that pays $10 if the coin comes up heads. I'll say that your doxastic state *condones* a bet when this is the case. We can conveniently represent a bet costing $n \times \$D$ that pays $\$D$ if *P* is true as follows:

Table 6: Bet on P

Bet on *P*	
P	$(1 - n) \times \$D$
¬*P*	$-n \times \$D$

Note that if this bet is condoned, then so is this one:

Table 7: Bet on ¬P

Bet on ¬*P*	
P	$-(1 - n) \times \$D$
¬*P*	$n \times \$D$

This is just the first bet, viewed from the other side. I'll say that cr(*P*) = *n* condones both these bets.[8]

---

resemble a Dutch Strategy. Accordingly, I do not adopt this terminology, and simply refer to all the arguments as DBAs.

[8] Note that this use of the term 'condone' is fitting. For example, if my credence that the coin will land heads is 0.4, I condone a bet that costs $4 and pays $10 if the coin lands heads, and I condone a bet that costs $6 and pays $10 if the coin lands tails.

With this connection between doxastic states and bets, we can demonstrate a flaw in a doxastic state by showing that the doxastic state condones a flawed set of bets. For instance, consider the synchronic constraint, *Finite Additivity*: $cr(A \vee B) = cr(A) + cr(B)$ for all A, B such that $(A \wedge B)$ is contradictory. Imagine that I have the following credences that violate *Finite Additivity*:

$cr(Heads) = 0.2$ $\quad$ $cr(\neg Heads) = 0.9$

One can demonstrate the flaw by offering me the following two bets, both of which my doxastic state condones:

Table 8: Bet Heads/¬Heads

Bet on *Heads*	
Heads	\$8
¬*Heads*	–\$2

Bet on ¬*Heads*	
¬*Heads*	\$1
Heads	–\$9

If *Heads* is true, then I lose \$1. If ¬*Heads* is true, I lose \$1. So, no matter what, I lose \$1. The thought is that something must have gone wrong with my doxastic state to condone bets which lead to guaranteed losses. Note that what we've established is the Dutch Book Theorem (with respect to *Finite Additivity*): If you violate *Finite Additivity*, then there exists a Dutch Book against you.[9]

So, how do we go from the Theorem to the Argument? Well, we could set up the argument as follows:

(1) If you violate *Finite Additivity*, then there exists a Dutch Book against you.

(2) If there exists a Dutch Book against you, then you are irrational.

(C1) Thus, if you violate *Finite Additivity*, then you are irrational.

(C2) Thus, you should obey *Finite Additivity*.

As Hájek ([2005]) notes, however, the move from (C1) to (C2) is suspect. For all the argument says, there could exist a Dutch Book against you if you obey *Finite Additivity*. If you're susceptible to guaranteed losses no matter whether you obey the principle or not, then we don't have an argument for obeying the principle. Thus, to fully complete the argument, we need the *Converse Dutch Book Theorem*. This says:

> If an agent obeys principle *p*, then there does not exist a set of bets such that the agent accepts all the bets and the bets make up a Dutch Book.

If we have the Converse Theorem, then we could make a case for obeying the principle in question. However, it is very difficult to prove the Converse Theorem. This is because, even if you obey some principle (like *Finite Additivity*), you might be vulnerable to a Dutch Book for some *other* transgression of a principle of rationality. Unless we state the One Unified Principle of Probabilistic Rationality all at once, we'll have trouble showing the Converse Theorem. Thus, it is often acceptable to show the *Weak Converse Dutch Book Theorem*:

> If an agent obeys principle *p*, then **possibly**[10] there does not exist a set of bets such that the agent accepts all the bets and the bets make up a Dutch Book.

If we have the Weak Converse Theorem, then we can still make the case for the principle in question: if you violate it you are definitely open to a Dutch Book; if you don't, then you may not be. When it comes to *Finite Additivity*, we can establish the Weak Converse Theorem. Thus, we can make an argument in favor of *Finite Additivity*, that looks something like this:

---

[9] Note, we haven't actually *proved* this, since I've only given one instance. But in seeing the one instance, one can verify to oneself that there is a proof.

(1) If you violate *Finite Additivity*, then necessarily there exists a Dutch Book against you.

(2) If there necessarily exists a Dutch Book against you, then you are irrational.

(C1) Thus, if you violate *Finite Additivity*, then you are irrational.

(3) If you obey *Finite Additivity*, then possibly there does not exist a Dutch Book against you.

(4) If possibly there does not exist a Dutch Book against you, then possibly you are not irrational.

(C2) Thus, if you obey *Finite Additivity*, then possibly you are not irrational.

(C3) Thus, you should obey *Finite Additivity*. [from (C1) and (C2)]

Now that I've introduced the basics of a Dutch Book Argument, I can note how our focus will be somewhat restricted. I've noted that before one can give a complete DBA for a principle, one must not only have the Dutch Book Theorem but also (at least) the Weak Converse Dutch Book Theorem. However, for the purposes of this chapter, I'll be focusing exclusively on the Dutch Book Theorem. This restriction is justified by the fact that having a Dutch Book Theorem for a principle is clearly a necessary part of having a complete Dutch Book Argument for that principle. So, in what follows I will focus on how we establish the Dutch Book Theorem for *COND*, *Self-Respect*, and *Reflection*.

[10] This should be understood as metaphysical possibility. We want to know if, in obeying *p*, there is some metaphysically possible way in which the agent can avoid condoning a set of bets that constitute a Dutch

Before presenting this, however, there is one important thing to note about the Dutch Book scenarios. I said that we reveal a problem with an agent's doxastic state by *offering* the agent the bets in question (above, the bets were the Bet on *Heads* and Bet on ¬*Heads*). But though actually offering the bets and having the agent accept them is a way to make vivid the error in the agent's doxastic state, the *actual* offering and accepting of bets is immaterial. What is important, and what is in error, is that the agent's doxastic state *condones* these two bets. This is true whether or not the bets are actually offered. In what follows, I'll use the phrase 'your doxastic state condones a bet' to emphasize that the actual offering and accepting of the bets is not crucial to successfully showing errors in the doxastic state. The error is already there, whether or not it is exploited by a clever bookie.[11]

6.3 Conditionalization Dutch Book

6.3.1 Briggs's Presentation

Since I will be focusing on Briggs's proposal about how to understand DBAs, I will present Briggs's version of the DBA for *COND*.[12] Since *COND* is a principle about what to do upon receiving evidence, we need to understand how this evidence

---

Book.

[11] Thinking about Dutch Books in this way also allows us to control for things we do not care about. For instance, there may be reasons unrelated to the features of the agent's doxastic state we are trying to assess, which would result in the agent not accepting a sequence of bets offered by a bookie, even though his doxastic state condones those bets. For instance, perhaps the bookie doesn't look trustworthy, or perhaps the agent doesn't have any interest in betting. If these situations obtained then the agent's credence function might condone bets even though he wouldn't actually accept them if they were offered them. We want to abstract away from these features, and focus purely on the agent's doxastic state. This picture of Dutch Book scenarios is clearly indebted to David Christensen's picture of them in his ([1996]).

[12] The *COND* DBA first appeared in Teller ([1973]), and is credited to Lewis. For Lewis's version see Lewis ([1999]) pp. 403-407.

acquisition is to go. According to Briggs, the following two assumptions about how this goes are sufficient to enable us to give a DBA for *COND*. First we need:

Partition Assumption: The agent's possible evidence at t_1 forms a partition, $\{E_i\}$. By this we mean that the disjunction of every member of $\{E_i\}$ is a tautology and the conjunction of any two members is contradictory. A simple situation where this obtains is where the evidence partition is $\{E, \neg E\}$. Second, we have:[13]

Veridical Assumption: For all members of $\{E_i\}$: (i) $\text{cr}(E_i) = 1 \rightarrow E_i$, and

(ii) $\text{cr}(E_i) = 0 \rightarrow \neg E_i$

Since we are considering agents that satisfy PROB, $\text{cr}(E) = 0$ iff $\text{cr}(\neg E) = 1$. Thus, the second condition in the Veridical Assumption is equivalent to

(ii′) $\text{cr}(\neg E_i) = 1 \rightarrow \neg E_i$.

Thus, the Veridical Assumption is a constraint that says that if you become fully confident in one of your evidence propositions, then it is true. Since for this chapter, I'll identify $\text{cr}(E) = 1$ with learning E, we can say that this assumption amounts to the assuming that if you learn E, then E is true.

Consider now an agent at *t0*, such that for some E in $\{E_i\}$:

$\text{cr}_0(A|E) = n$

$\text{cr}_0(E) = d \ \ (0 < d < 1)$

$\text{cr}^E(A) = r \ (r \neq n)$

Let 'cr^E' refer to the credence function upon learning E and only E. This agent is set up to violate *COND*: if at *t1* he learns E, he will violate that principle.[14] We can display

[13] Briggs says we must assume "…that that she has no chance of mistaking her evidence—that is, if $Cr(E) = 1$ after she updates, then E is true, and if $Cr(E) = 0$ after she updates, then E is false." (p. 62)

[14] Note that the Veridical Assumption and the Partition Assumption together imply that if the agent learns E, then he learns *only* E. According to condition (i): $\text{cr}(E) = 1 \rightarrow E$. So imagine that the agent learns both

what is wrong with such a doxastic state by considering the following series of bets. At *t0* we offer:

Table 9: Bets C1 and C2

Bet C1 on $A\|E$	
$A \wedge E$	$1 - n$
$\neg A \wedge E$	$-n$
$\neg E$	0

Bet C2 on E	
E	$(d-1)(r-n)$
$\neg E$	$d(r-n)$

Given the agent's credences Bets C1 and C2 are condoned.[15] At *t1* the agent either learns *E*, or he doesn't. If he doesn't, we do nothing. If he *does* learn *E* we offer him:

Table 10: Bet C3

Bet C3 on A	
A	$r - 1$
$\neg A$	r

which is condoned given that $cr^E(A) = r$. Given this strategy, it is alleged, the agent is certain to suffer a loss. We show this by considering two cases:

> $\neg E$ is true: If $\neg E$ is true, then (via the Veridical Assumption) the agent doesn't learn *E*, so Bet C3 is off. Bet C1 yields 0 and Bet C2 yields $d(r - n)$.
>
> E is true: If *E* is true, then the agent learns *E* so all three bets are on. The agent wins a total of $(r - n)$ on Bets C1 and C3. On Bet C2, the agent wins $(d - 1)(r - n)$. Adding these winnings together we get $d(r - n)$.

If $r < n$, then we offer the agent these bets, and his total winnings are negative. If $r > n$ then we offer the agent the other side of these bets (on $\neg A|E$, $\neg E$, and $\neg A$), and his total winnings are negative. So, if $r \neq n$, the agent loses. If the agent is set up to violate

---

E_1 and E_2. The Veridical Assumption entails that both E_1 and E_2 are true. But this violates the Partition Assumption, which states that E_1 and E_2 are contradictory.

[15] Bet C1 is a conditional bet. A conditional bet on $A|E$ is just like a normal bet on *A*, except that the bet is voided unless *E* is true. Conditional bets are condoned by conditional credences in a natural way: if

COND, then there is some distribution of credences where $r \neq n$. So, if the agent is set up to violate *COND*, then he is open to a set of bets guaranteed to lose money. This is Briggs's presentation of the *COND* DBA.

6.3.2 The Learning Assumption

There is, however, an immediate problem with this presentation. In considering the case where E is true, I assumed that if E is true then the agent learns E. But why suppose this? It isn't entailed by the Veridical Assumption, which gives us the converse, and the Partition Assumption doesn't provide this either. So, we haven't ruled out the scenario where E is true, and yet it isn't learned. In such a scenario Bet C3 is never made, so we only consider Bets C1 and C2, yielding two scenarios:

E true:	$\neg E$ true:
$(1 - n) + (d - 1)(r - n)$	$-n + (d - 1)(r - n)$
$=$	$=$
$1 + [d(r - n) - r]$	$d(r - n) - r$

But here we do not have guaranteed losses.

To fix the *COND* DBA, then, we'll have to make another assumption. The assumption we need is:

Learning Assumption: $E_i \rightarrow E_i$ is learned

We retain the Partition Assumption, but jettison the Veridical Assumption. This is because the Learning Assumption (LA) and the Partition Assumption (PA) entail the Veridical Assumption. According to PA, if $\neg E_k$ is true, then there is some $E_{i \neq k}$ that is

---

$\mathrm{cr}(A|E) = n$ then this condones a bet on $A|E$ that costs $n \times \$D$, pays $\$D$ if $A \wedge E$ is true, and returns the cost of the bet if E ends up being false.

true. According to LA, this E_i is learned. But if E_i is learned, then E_k is not learned, for if the agent were such that $cr(E_i) = 1$ and $cr(E_k) = 1$, then he would violate PROB. Thus, we have that if $\neg E_k$ is true, then E_k is not learned. This is the Veridical Assumption.

These two assumptions, LA and PA, handle the problem case for the *COND* DBA. The problem, recall, resulted from the fact that *E* could be true, and yet the agent not learn it. LA rules this out. Thus, with LA and PA as our assumptions, Bets C1-C3 offered in the way indicated guarantee losses for any violator of *COND*.

6.3.3 The Reflection and Self-Respect DBAs

The DBA for *Reflection* is almost identical to the one for *COND*. Consider the following *Reflection*-violating credence function:

$cr_0(A|\langle cr_1(A) = r\rangle) = n \ (r \neq n)$

$cr_0(\langle cr_1(A) = r\rangle) = d \ (0 < d < 1)$

To display why this is problematic, we offer the agent with such credences a series of bets, structurally analogous to those above:

Table 11: Bets R1 and R2

Bet R1	
$A \wedge cr_1(A) = r$	$1 - n$
$\neg A \wedge cr_1(A) = r$	$-n$
$cr_1(A) \neq r$	0

Bet R2	
$cr_1(A) = r$	$(d - 1)(r - n)$
$cr_1(A) \neq r$	$d(r - n)$

Table 12: Bet R3

Bet R3	
A	$r - 1$
$\neg A$	r

Bets R1 and R2 are made at *t0*. Bet R3 is made at *t1* iff $cr_1(A) = r$. Consider, now, the two scenarios:

$\underline{cr_1(A) \neq r}$: If this is the case, then Bet R3 is called off, and Bet R1 pays 0. Thus, the agent wins $d(r - n)$.

$\underline{cr_1(A) = r}$: If this is the case, then Bet R3 is on. The agent wins a total of $(r - n)$ on Bets R1 and R3. On Bet R2, the agent wins $(d - 1)(r - n)$. Adding these winnings together we get $d(r - n)$.

Just as in the *COND* DBA, so long as $r \neq n$, we can set things up so that the agent is guaranteed to lose. Since if one fails to satisfy *Reflection* there is some such distribution of credences where $r \neq n$, it follows that if the agent violates *Reflection*, then he condones a set of bets that guarantee a loss.

The DBA for *Self-Respect* is very similar. In the presentation above, simply make every time index the same. For convenience, I will refer to the bets in the *Self-Respect* and *Reflection* DBAs as 'R1', 'R2', and 'R3'. But one should note that strictly speaking, the bets in the *Self-Respect* DBA would need to be changed, to reflect the change in the time indices. Context should make clear which bets I am referring to later in the paper.

Note that in the *Reflection*/*Self-Respect* DBAs, there is no need for PA or LA. We do not need PA, since these DBAs do not depend on the agent getting evidence in any way, and PA is about the structure of this evidence. LA stipulates that a certain proposition, if true, will be learned. Again, since these DBAs do not concern learning experiences, we do not need LA.

6.4 Christensen/Briggs Response to the Reflection/Self-Respect DBAs

Christensen holds that neither *Reflection* nor *Self-Respect* are requirements of rationality. Thus, he needs a response to the DBAs that seem to establish these norms. Here is Christensen's response to the *Self-Respect* DBA:

> In [standard Dutch Book arguments], the set of bets is one whose payoffs are logically guaranteed to leave me poorer. Not so for the bets involved in the [*Self-Respect* Dutch Book]. Consider the case where I don't actually have [*r*] credence in A. Here, the bookie does not offer me Bet 3, so the relevant set of bets is just 1 and 2. But this set is not one whose payoffs logically guarantee my loss. True, they'll cost me money in the actual world. But in a world where I do have credence [*r*] in A, and where A is false, I win… ([2007], p. 329)

Briggs claims that *Reflection* is not a norm of rationality (she doesn't discuss *Self-Respect*). She gives a response similar to Christensen in objecting to the *Reflection* DBA:

> …the Dutch book against agents who violate *Reflection* reveals diachronic self-doubt. At a world where the agent makes bets 1 and 2 […], he or she is already guaranteed to suffer a net loss. But as long as the agent doesn't make bet 3, there are counterfactual worlds where he or she enjoys a net gain. At those counterfactual worlds, of course, the agent's beliefs would have condoned different betting behavior. But the bets we consider at counterfactual worlds are fixed by the agent's actual (not counterfactual) credence function. ([2009], p. 82)

These responses should initially strike one as odd. The claim being made is that the set of bets do *not* guarantee a loss if one were to have a doxastic state that didn't condone Bet R3 but allowed one to receive the payouts from Bets R1 and R2 as if one *did* have such a doxastic state. Although this is true, it isn't obvious how it helps: there *isn't* such a doxastic state! If your doxastic state doesn't condone Bet R3, then you can't collect winnings on Bets R1 and R2 that depend on having a doxastic state that does.

According to the response that Christensen and Briggs give, however, something like this is true. So just what is the idea? First of all, it is certainly true that any world in

which R3 is not offered or condoned, the sum of the payouts on R1 and R2 are negative. Further, it is certainly true that any world in which R3 is offered and condoned is a world in which the sum of the payouts on R1, R2, and R3 are negative. So, there is a perfectly good sense according to which the agent is guaranteed to lose money in the *Reflection*/*Self-Respect* DBA. No one is denying these facts. The claim being made here by Christensen and Briggs is that there are different ways in which a set of bets guarantee a loss of money, and not all of these ways are indicative of irrationality in the agent's doxastic state.

For instance, according to Christensen, the guaranteed losses incurred by a set of bets only tell against the agent's doxastic state if that set of bets guarantees a loss *in every possible world*. Now, it is true that R1, R2, and R3 guarantee a loss in every possible world *in which they are condoned*. But, according to Christensen, this is not enough. They must guarantee a loss in every possible world. But they do not do this. Consider a world where $cr_0(A) \neq r$ and $\neg A$ is true. This is a possible world. In that world R1 pays 0, R2 pays $d(r - n)$, and R3 pays r. So, the total payout is $d(r - n) + r$. There are values of d, r, and n with $r \neq n$ where this is positive (for instance: $d = 0.5$, $r = 0.6$, $n = 0.8$). Accordingly, R1, R2, and R3 do not result in a loss in *every possible world*. Thus, according to Christensen, the *Self-Respect* DBA doesn't show the right kind of guaranteed losses and so fails.

Why would someone hold a view like this? At one point Christensen says that this is just how Dutch Books work: we must show losses at *all possible worlds*. But a more substantive reason is suggested by his discussion. The idea seems to be that when we have to show losses in *all possible worlds*, this guards against the possibility that the

betting losses are simply due to ignorance rather than irrationality. Consider a simple situation where I can guarantee that you actually lose money: I know that the Red Sox lost, but you do not. Thus, I can offer you a bet—at odds you deem fair—that you win if the Red Sox won and lose otherwise. You are certain to lose money on this bet, but this shows no irrationality in your doxastic state. Your ignorance explains your guaranteed loss, not irrationality. Now, one way to make sure that the betting losses are not due to ignorance, is to evaluate the payouts of bets at *all possible worlds*. If we did this in the example just given, then we get the right result: the bet I offered you doesn't lose in all possible worlds (in particular, not in worlds where the Red Sox won), and so the guaranteed losses don't reveal irrationality in your doxastic state.

Briggs adopts a response to the *Reflection* DBA very similar to Christensen's response to the *Self-Respect* DBA, but the reason behind her response is different. She argues that there are two importantly different roles that an agent's doxastic state plays in the *Reflection* DBA. First, the agent's doxastic state determines how confident she is in various propositions, and so determines which bets she condones. Secondly, the agent's doxastic state determines the payouts of certain bets.[16] For example, in the *Reflection* DBA $\mathrm{cr}_1(A) = r$ plays two roles. It condones bet R3 on A, but is also the target of Bet R2 in that the truth-value of $\mathrm{cr}_1(A) = r$ determines the payout of R2.

Briggs's idea is that if an agent's doxastic state operates in the bet-condoning role in an incoherent way, then we have a kind of rationally objectionable incoherence. It is objectionable because the bet-condoning role that your doxastic state plays in a

[16] Briggs writes:

> An agent's belief function fixes the truth values of her beliefs in two ways. First—almost tautologically—it fixes what she believes. Second and less obviously, it fixes the truth values of some of her higher-order beliefs. ([2009], p. 80)

DBA corresponds precisely to the role that your doxastic state plays in fixing what you believe.[17] If there is some incoherence here, the thought goes, then there is an objectionable kind of incoherence in what you believe. But, Briggs notes, this is not what is happening in the *Reflection* DBA. In the *Reflection* DBA the agent loses in all worlds in which the bets are condoned as a result of the fact that $cr_1(A) = r$ plays two roles that are leveraged against one another. Since the agent is betting *according to* the function 'cr_1' and betting *on* certain features of the function 'cr_1', R2 is lost *iff* R3 is condoned. But, the thought goes, these kinds of losses don't home in on incoherence in the bet-condoning role of the agent's doxastic state.[18] So, the *Reflection* DBA fails for a similar reason that the *Self-Respect* DBA fails. It shows that the agent loses at all worlds where the bets are condoned, but not that the bets lose at all the worlds that matter (for more technical discussion of the mechanics of this response, see Appendix A).

Christensen is clear that for us to have a successful DBA, the bets under consideration must result in losses at *all* possible worlds. Briggs is less explicit about which worlds must be worlds where there are losses. Despite this, she is at least committed to the claim that we must evaluate the payouts of bets at more worlds than just those where the bets are condoned. Further, for her response to work against the *Reflection* DBA, she must be committed to the claim that the payouts of bets R1, R2, and R3 are considered at worlds where $cr_1(A) \neq r$. This ensures that there are some

[17] This is clear from the way we set up the relation between doxastic states an bets condoned. If your doxastic state is such that $cr(P) = n$, then this determines that you believe P to degree n. Similarly, if your doxastic state is such that $cr(P) = n$, then this condones a bet on P proportional to n.

[18] Briggs writes:

> In cases of incoherence, the agent is guaranteed to be wrong solely because of how her beliefs operate in the first, belief-fixing role. In Moore's paradox cases and cases of self-doubt, the agent is guaranteed to be wrong because of some faulty interaction between the two roles. ([2009], p. 80)

worlds where R1, R2, and R3 do not result in losses. So, it may be that Briggs agrees with Christensen that the kinds of guaranteed losses that matter are losses that occur at *all* possible worlds. But if not, then at the very least, Briggs must be committed to the following claim:

> A set of bets (like R1, R2, and R3) do not guarantee losses in the sense that matters if those bets win in worlds where a proposition playing two roles in those bets (like $cr_1(A) = r$) has a truth-value different than its truth-value when those bets are condoned.

This ensures us that whenever a proposition plays two roles, we can isolate the bet-condoning role as Briggs desires.

There are important questions to ask about Christensen's response and Briggs's response. With respect to Christensen we can ask if we really need to have losses in *all possible worlds* to ensure that we are not taking advantage of the agent's ignorance. With respect to Briggs we can ask whether it is really objectionable to leverage two roles that a proposition can play against the agent. The proposal that I will give attempts to provide answers to these questions. However, before getting to this proposal, I first will explain how this kind of response to the *Reflection*/*Self-Respect* DBAs requires one to reject the *COND* DBA. This will show that Briggs's response to the *Reflection*/*Self-Respect* DBA does not allow us to distinguish the *COND* DBA from the *Reflectin*/*Self-Respect* DBA as Briggs claims it will.

6.5 Analogous Response to the Conditionalization DBA

Briggs and Christensen reject the *Reflection*/*Self-Respect* DBA because the Dutch Book Theorem is not established: violation of *Reflection* and *Self-Respect* do not lead to

losses at all the worlds that matter. Note that an analogous response is not available when it comes to the DBA for an agent that violates *Finite Additivity*. The bets offered to such an agent result in losses at *all possible worlds*, so it is trivial that the agent loses in all the worlds that matter. Thus, this response allows us to distinguish the *Reflection*/*Self-Respect* DBAs from the DBA for *Finite Additivity*.

Nevertheless, the response, if successful, shows that the *COND* DBA is flawed, too, and so doesn't allow us to distinguish the arguments as Briggs claims. According to the *COND* DBA setup, the set of bets condoned is either C1, C2, and C3, or just C1, and C2. Focus on this latter scenario, where just C1 and C2 are offered and condoned. In such a scenario, the agent fails to learn E (since it is failure to learn E that determines that C3 is not offered). Given that the agent fails to learn E in this scenario, it follows from LA that E is false in these scenarios. In all the worlds where only C1 and C2 are condoned these two bets result in guaranteed losses of $d(r - n)$. But if we adopt the response to the *Reflection*/*Self-Respect* DBA above, then we must consider more than just these worlds when evaluating the payouts of the two bets.

If we adopt Christensen's proposal we must consider the payout of these two bets at *all possible worlds*. There is a possible world where E and A are both true. In such a world, C1 and C2 together result in $(1 - n) + (d - 1)(r - n)$, which is positive for some values of d, r, and n. Thus, the *COND* DBA is not one where we have losses at *all worlds*. It is true, of course, that to condone *only* C1 and C2, the agent must fail to learn E. And it is also true that if the agent fails to learn E, then E is false. So, if the agent were to actually be offered (and accept) only C1 and C2, he would lose money. But this is irrelevant given the proposal articulated above. For in the *Reflection* DBA, to

condone *only* R1 and R2, the agent must be such that $cr_1(A) \neq r$. And if $cr_1(A) \neq r$, then R1 and R2 result in a loss of money. So, if the agent were to actually be offered (and accept) only R1 and R2, he would lose money. But this is alleged to be of no importance in the *Reflection* DBA, so it must be of no importance in the *COND* DBA. Thus, if a good DBA is one that shows losses in *all worlds*, then we can't distinguish the *Reflection*/*Self-Respect* DBAs from the *COND* DBA.[19] Both are flawed.

Above I noted that Briggs isn't clear about whether we need to have losses at *all* worlds for a DBA to be convincing or just at some worlds. However, I noted that she must at least be committed to the following claim:

> A set of bets (like R1, R2, and R3) do not guarantee losses in the sense that matters if those bets win in worlds where a proposition playing two roles in those bets (like $cr_1(A) = r$) has a truth-value different than its truth-value when those bets are condoned.

One might think that if Briggs is only committed to this, then we can still distinguish the *COND* DBA from the *Reflection* DBA. The idea is that although in a world where *E* and *A* are true C1 and C2 do not result in losses, this doesn't matter. *E* is false at all worlds where only C1 and C2 are condoned and since it *doesn't* play two roles in the bets, we don't need to worry about worlds where *E* is true. In this way, the thought goes, it is different than $cr_1(A) = r$. This claim, however, cannot be maintained. For *E* *does* play two roles in the *COND* DBA. *E*'s truth determines whether *E* is learned and so affects whether or not Bet C3 is condoned. But in addition, *E*'s truth determines the payout of Bet C2.

[19] In his [2007] Christensen has a footnote suggesting this line of argument as a way of criticizing the *COND* DBA.

Here, then, are how things stand. Christensen and Briggs have independently proposed a response to the *Reflection*/*Self-Respect* DBAs. Both are proposals for how to understand what kinds of guaranteed losses matter for DBAs. According to Christensen, a set of bets results in guaranteed losses in the sense that matters if and only if the bets lose in *all possible worlds*. Briggs may assent to this, too, or she may make the weaker claim that a set of bets results in guaranteed losses in the sense that matters if and only if the bets lose in *all worlds where the bets are condoned and in worlds where the truth-value of a proposition that plays two roles is different than its value when the bets are condoned*. We've seen, however, that no matter which claim one goes with, the *COND* DBA falls along with the *Reflection*/*Self-Respect* DBA. In what follows I will outline a proposal for how we can distinguish these arguments from each other.

6.6 The Evaluation World Proposal

6.6.1 The Guiding Idea

The proposal that I will outline builds off two main ideas. The first is the idea proposed by both Christensen and Briggs that not all guaranteed betting losses are created equal: only certain kinds of guaranteed betting losses matter when it comes to establishing the Dutch Book Theorem. The second idea is a familiar one, that was mentioned above: if a DBA is to be convincing, the guaranteed losses can't depend solely on the agent's ignorance. The proposal that I will give tells us that for a set of bets to reveal a doxastic flaw, the guaranteed losses must occur at more worlds than just the worlds at which the bets are condoned. However, my proposal will tell us how to pick out the relevant

worlds in a principled way that is meant to control for the agent's ignorance. Before I can explain how this works, we must have in hand some technical machinery.

6.6.2 Technical Machinery

Dutch Book Arguments start out with a strategy for offering bets. In the DBA for *Finite Additivity* this strategy is a simple one: a set of bets are offered at one time no matter what else happens to be true. In the DBAs for *COND*, *Reflection*, and *Self-Respect*, they are slightly more complicated. One set of bets is considered if one situation obtains, and another set of bets are considered if a different situation obtains.

A betting strategy can be thought of as a set of instructions. Formally, this is a function from worlds to sets of bets. If we let 'C1' indicate that bet C1 is offered, and let '$\underline{\text{C1}}$' indicate that C1 is not offered, then the strategy for the *COND* DBA is the following function:

Table 13: The COND DBA Strategy

{*w*: it is not the case that *E* is learned in *w*}	→	{C1, C2, $\underline{\text{C3}}$}
{*w*: *E* is learned in *w*}	→	{C1, C2, C3}

It is natural to think of this strategy as having two parts. There is the {C1, C2, $\underline{\text{C3}}$} part, and there is the {C1, C2, C3} part.

Now we need some terminology. Say that a world is a **condoning world** (**c-world**) relative to a part of a betting strategy. If we say that bet $\underline{\text{C3}}$ is condoned iff C3 is not condoned, then a world is a condoning world for a part of a strategy just in case in that world, every bet in that part of the strategy is condoned. Good Dutch Books are set up so that in every world assigned to a part of a strategy the bets for that part of the strategy are condoned. For example, an agent's doxastic state condones {C1, C2, C3}

just in case $cr_0(A|E) = n$, $cr_0(E) = d$, and $cr_1(A) = r$. Similarly, given the set up, an agent's doxastic state condones {C1, C2, C3} just in case $cr_0(A|E) = n$ and $cr_0(E) = d$ and $cr_1(A) \neq r$.

On the proposal that I will give, we are to evaluate the payouts of a set of bets not solely at the worlds where these bets would be condoned. To be clear about this, I introduce the notion of an **evaluation world**. A world is an evaluation world *relative* to a c-world. It is thus natural to think of evaluation worlds as determined by a privileged accessibility relation among possible worlds. In particular, the accessibility relation will be from c-worlds to evaluation worlds. Call this relation 'evaluation accessibility'.

The framework just presented is general enough to allow us to state the orthodox way of thinking about the payouts of bets. According to the orthodox way of evaluating the payout of a set of bets, we simply see what those bets yield in the worlds in which they would be condoned. If in all these worlds, the bets yield losses, then we say that the bets guarantee a loss of money. In the framework I've set up, this orthodox view holds that the evaluation accessibility relation is just the identity relation.

We can also state Christensen's view in this framework. According to Christensen, a set of bets results in guaranteed losses—in the sense that matters for DBAs—just in case in *all* worlds, those bets lose. Thus, according to Christensen *every* world is evaluation accessible from any c-world.

Briggs is somewhat less clear about what evaluation accessibility might be, but it seems to turn on whether or not some proposition plays two roles in the Dutch Book Argument. Suppose that some proposition, P, plays two roles in the bets offered. Now consider a c-world for those bets, c. Her view seems to be that the evaluation

accessibility relation at least guarantees us that a world where *P* is true, is evaluation accessible from *c*, and so is a world where ¬*P* is true.

Given this framework, we can now state more precisely why the Christensen/Briggs response requires us to reject the *COND* DBA. Consider Christensen's view first. Imagine a c-world for {C1, C2, C3}. This is a world where $cr_0(A|E) = n$, $cr_0(E) = d$, and ¬*E* is true. In this c-world, {C1, C2, C3} result in a loss. However, given Christensen's view about evaluation accessibility, there is an evaluation world for this c-world where *E* is true. In such a world, {C1, C2, C3} result in a gain. Thus, the part of the strategy characterized by {C1, C2, C3} does not result in guaranteed losses. Accordingly, the total strategy does not result in guaranteed losses. Consider now Briggs's view. Imagine again a c-world for {C1, C2, C3}. In this c-world, {C1, C2, C3} result in a loss. However, ¬*E* is playing two roles in this c-world. It both determines that C3 is not offered (via LA), and it is the target of C2. Accordingly, there must be an evaluation accessible world where *E* is true and where ¬*E* is true. But in a world where ¬*E* is true, as we just saw,{C1, C2, C3} result in a gain. So if there is an accessible evaluation world where ¬*E* is true, then this part of the strategy does not result in guaranteed losses. Accordingly, the total strategy does not result in guaranteed losses.

The proposal that I will give tells us something about the evaluation accessibility relation, but it is different than Briggs's view, Christensen's view, or the orthodox view. My proposal is easiest to state in terms of which worlds are *not* evaluation accessible from a given acceptance world.

Evaluation Worlds (EW): A world *e* is evaluation inaccessible from c-world, *c*, iff

(i) cr(*P*) = 1 in *c* at some time over which the bet strategy ranges,

(ii) *P* is true in *c*, and

(iii) ¬*P* is true in *e*.

Suppose we say that an agent *knows P* in world *w,* when cr(*P*) = 1 and *P* is true in *w*. Then, EW says that the evaluation worlds for a c-world are those worlds consistent with what the agent knows in the c-world.

Given the framework I've proposed, a world is an evaluation world relative to a c-world. However, since each c-world is associated with one, and only one part of a betting strategy, it makes sense to say that a world is an evaluation world relative to a part of a betting strategy. In particular, we can say that a world is an evaluation world for a part of a betting strategy {B1, B2, …, B*n*} just in case that world is evaluation accessible from at least one c-world for {B1, B2, …, B*n*}. So, in evaluating the payout from a part of a betting strategy, we hold fixed the truth-value of any proposition that is true in all evaluation worlds for that set of bets. We let the truth-values of all other propositions vary.

To finish up the proposal, EW must be integrated into a condition on having a good DBA. I state this as a necessary condition, since I am focused solely on the part of the DBA involving the Dutch Book Theorem. Further, as I discuss later in this chapter, there may be other features of Dutch Book Arguments that cause them to fail. The following necessary condition is meant to take care of the kinds of cases that Christensen and Briggs focus on:

We have a good DBA for principle *p only if*

(i) there is a strategy for offering bets where each world where *p* is violated is assigned to a set of bets, and this assignment partitions the space of worlds,

(ii) the agent's credence function condones each of those bets at the c-worlds at the appropriate time, and

(iii) the set of bets leave the agent with a loss in all the evaluation worlds for that part of the strategy as determined by EW.

I have now presented my proposal. There are three things left to do. First, I will show that according to EW the *COND* DBA results in guaranteed losses (of the sort that matter), whereas the *Reflection*/*Self-Respect* DBAs do not. Second, I will attempt to motivate EW. Third, I will respond to objections to this way of classifying DBAs.

6.7 EW in Action

We can see how the EW proposal performs by considering the *COND* DBA. According to the EW proposal, each set of bets has an associated set of evaluation worlds. So, each part of the strategy will have an associated set of evaluation worlds. For example, let's say that we want to know what the evaluation worlds are for {C1, C2, C3} and a particular agent, *S*. We proceed as follows. First, we list all the worlds in which that part of the strategy's bets would be offered. If conditions (i) and (ii) are met, these are just the c-worlds for this part of the strategy. Thus, in all these worlds, *E* is true. Next, we look at *S*'s doxastic state in all those c-worlds. In all these worlds, according to LA, $cr_1(E) = 1$. Thus, the set {C1, C2, C3} will have as evaluation worlds all and only the worlds where *E* is true. All other propositions are true in some of the evaluation worlds

for that set of bets and false in others. Accordingly, this part of the strategy is guaranteed to result in a payout of $d(r - n)$.

Consider the other part of the strategy, which is characterized by the set {C1, C2, C3}. In all the c-worlds for this part of the strategy, the agent fails to learn *E*. This is what determines that C3 is not part of the strategy. According to LA, if the agent fails to learn *E*, then *E* is false. So all the c-worlds for this part of the strategy are world where *E* is false. But, according to PA, one of the $\{E_i\}$ is true in each c-world, and so from LA it follows that for some $E_i \neq E$, $\mathrm{cr}_1(E_i) = 1$. Thus, every evaluation world for this part of the strategy is one where some $E_i \neq E$ is true. Since every $E_i \neq E$ is inconsistent with *E*, these are all worlds where *E* is false. Thus, all the evaluation worlds for {C1, C2, C3} are worlds where *E* is false. Accordingly, this part of the strategy results in a payout of $d(r - n)$. When we put this together with the first part of the strategy, we see that at *all* evaluation worlds for this strategy the payout is $d(r - n)$. Thus, the *COND* DBA bets guarantee a loss in the sense that matters for DBAs.

The *Reflection/Self-Respect* DBAs fare differently. Consider the *Self-Respect* DBA first. It can be constructed so as to meet conditions (i) and (ii). It meets condition (i), because there are two parts of the strategy for offering bets, one if $\mathrm{cr}(A) \neq r$ and one if $\mathrm{cr}(A) = r$. Since these propositions partition all the worlds, condition (i) is met. However, the *Self-Respect* DBA fails with respect to (iii). Consider the part of the strategy characterized by {R1, R2, R3}. It is essential to Bet R2 of the *Self-Respect* DBA that $0 < \mathrm{cr}_0(\langle \mathrm{cr}_0(A) = r \rangle) < 1$. Thus, the c-worlds for this part of the strategy are worlds where $0 < \mathrm{cr}_0(\langle \mathrm{cr}_0(A) = r \rangle) < 1$. Given this, EW says that there are evaluation worlds for this set of bets where $\mathrm{cr}_0(A) = r$ and evaluation worlds where $\mathrm{cr}_0(A) \neq r$. But

if there are evaluation worlds for {R1, R2, R3} where $cr_0(A) = r$, then this part of the strategy does not guarantee losses in the sense that matters. For in worlds where $cr_0(A) = r$, R1 and R2 both win. Thus, we don't have the kind of guaranteed loss required.

In a similar way, the *Reflection* DBA fails. The *Reflection* DBA meets conditions (i) and (ii), however it too fails condition (iii). Consider again the part of the strategy characterized by {R1, R2, R3}. Just as above, it is critical to the *Reflection* DBA that at *t0*, $0 < cr_0(\langle cr_1(A) = r\rangle) < 1$. So, at all c-worlds $0 < cr_0(\langle cr_1(A) = r\rangle) < 1$. Further, from the fact that the agent has *Reflection*-violating credences, nothing follows about the value of $cr_1(\langle cr_1(A) = r\rangle)$. So, there are c-worlds where at $0 < cr_1(\langle cr_1(A) = r\rangle) < 1$. Thus, there are many c-worlds where the agent never has the information that $cr_1(A) = r$. Given this, there will be evaluation worlds for this set of bets where $cr_1(A) = r$ and worlds where $cr_1(A) \neq r$. But in a world where $cr_1(A) = r$, R1 and R2 both win. So just as for the *Self-Respect* DBA, the *Reflection* DBA betting strategy does not guarantee a loss in the sense that matters.

6.8 Motivating the EW Proposal

I have now sketched a framework for thinking about DBAs and a proposal that gives the result that the *Reflection*/*Self-Respect* DBAs are flawed because they do not display guaranteed losses in the sense that matters. Nevertheless, my proposal allows us to criticize these arguments in a way that does not carry over to the *COND* DBA. One might think that despite this, there is no motivation for understanding DBAs in this way.

One response to this challenge is to point out that this way of understanding DBAs is motivated by the fact that it gives the result that the DBAs for *COND* and

Finite Additivity[20] are good and that the DBAs for *Reflection* and *Self-Respect* are no good. This does provide some motivation for this way of understanding DBAs. It is a piece of data that *Reflection* and *Self-Respect* are unintuitive while *COND* and *Finite Additivity* are intuitive. A way of understanding DBAs that makes the arguments for the intuitive principles come out as good, and the arguments for the unintuitive principles come out as no good receives some motivation from this very fact.

But this is less motivation than we'd ultimately like. We might be legitimately confused about whether or not *COND* is a normative principle. The *COND* DBA does little to shine a light on this issue if we're told that it is a good argument, according to a way of classifying DBAs that receives its sole motivation from the fact that it classifies the *COND* DBA as good. So, we need some motivation for this way of understanding DBAs.

I believe that one can provide motivation for this proposal. Recall that EW states the evaluation worlds for a c-world are those worlds consistent with what the agent knows in the c-world. The guiding idea behind EW is that it is a way of controlling for the ignorance of the agent whose doxastic state we are evaluating. This is a familiar idea when it comes to DBAs. The idea is that guaranteed losses don't mean much if they are merely the result of some ignorance on behalf of the agent. I'm guaranteed to lose a bet at 1:1 odds on *Heads* if unbeknownst to me the coin has already landed *Tails*. But this shows no irrationality. The motivation for EW is that it is a particularly clear way of controlling against taking advantage of ignorance in this way. If a set of bets guarantee losses according to EW, then we can be sure that the losses are not due to ignorance.

[20] The DBA for *Finite Additivity* is counted as good because the agent is guaranteed to lose at *all worlds* in that DBA. Since the evaluation worlds for that strategy are a subset of all the worlds, the agent is

In standard presentations of DBAs, one controls for this by claiming that a Dutch Book is flawed if the bookie requires more information than the bettor to make the book. For three reasons, this is an unattractive way of thinking about things. First, it simply isn't clear *what* information the bookie needs to make a book. Does the bookie, for instance, need to know the agent's policy for updating to implement the *COND* book? Some think *yes*.[21] But there's a sense in which the bookie doesn't need to know this. If the agent is set up so as to violate *COND*, the bookie can luckily offer the agent Bets C1 and C2 without knowing the agent's update policy. If the agent is not a Conditionalizer in the sense that $\mathrm{cr}^{E}(A) \neq r$, then he has locked himself into a set of bets that guarantee a loss. This seems to show just as much a flaw in the agent's doxastic state as the situation where the bookie knows the update policy. So, perhaps the bookie need not have this information.

There is a second reason why controlling for losses due solely to ignorance should not be done by maintaining that the bookie and bettor have the *same* information. Above I explained how the bookie and betting situation is a bit of fiction meant to make the scenario more vivid. There doesn't actually have to be a bookie in the Dutch Book scenario. But if there isn't actually a bookie then it doesn't make sense to ask whether the agent and the bookie have the same information.

But even waiving this objection there is a problem. Presumably if the bookie is to actually be able to Dutch Book the bettor, the bookie must know that the bets he is about to offer are guaranteed to make the bettor lose money. But if it is a general rule of Dutch Books that the bookie and bettor must have the same information, then we must

---

guaranteed to lose at all evaluation worlds.

[21] For instance, Vineberg ([1997]).

give the bettor the information that the bets he is about to accept are guaranteed to make him lose money. But if we give the bettor this information, then there is no reason to think that the bettor would accept the bets.

For these reasons, EW gives us a better way of controlling for the agent's ignorance. It doesn't matter *what* information a bookie would need to set up the book. Instead, we evaluate the bets without keeping fixed information the agent doesn't have. Doing this ensures that the agent's ignorance alone is not what guarantees the betting losses. This allows us to bypass worries about whether or not a non-essential bookie has the same information as the agent.

This is the general motivation for EW. But more specific motivation can be give. EW is a bi-conditional. We can consider each direction of the bi-conditional in turn:

> **EW.1:** If (in c) there is no time over which the bet strategy ranges where the agent knows that P, then there is at least one evaluation world (relative to c) where P is true, and one where P is false.

> **EW.2:** If the agent knows that P (in c) at some time over which the bet strategy ranges, then no world e where P is false is an evaluation world (relative to c).

Consider EW.1 first. Note that there are two ways in which the agent can fail to know that P: either $\text{cr}(P) \neq 1$ or $\text{cr}(P) = 1$, but P is false. Consider the former case. Here the agent thinks that $\neg P$ is possible. But we shouldn't take evaluation worlds out of play that the agent thinks are possible. For if we were to do such a thing, then we might just

be punishing the agent for a lack of information, rather than a lack of coherence between her doxastic states. To protect against this, we allow evaluation worlds where P is true and evaluation worlds where P is false. Consider now the latter case. Here the agent thinks that $\neg P$ is impossible. Perhaps, the thought goes, it is acceptable to have no evaluation worlds where $\neg P$ is true. However, if we do this then we do not allow the c-world itself to be an evaluation world. This is highly counterintuitive since the world in which the agent condones the bets surely is relevant to the evaluation of those bets' payouts. Thus, EW.1 allows us to control correctly for the agent's ignorance. If at all evaluation worlds we can show a loss, then we know this is not due simply to the agent lacking some information.[22]

Next consider EW.2. Note first that it is plausible to think that if an agent is completely certain that a proposition is true, and if he's right about this, then he is criticizable for betting in such a way that gain is possible only if that proposition is false. So, EW.2 can be motivated in the following way: we permit this kind of criticism by having no evaluation worlds where a known proposition is false. This is where my proposal differs from Christensen's. Christensen has every world as an evaluation world for any c-world. This controls for the agent's ignorance in much the way that EW.1 does. However, my claim is that we do not need to allow *all* worlds as evaluation worlds for any c-world. In fact, we do not *want* to allow all worlds as evaluation worlds. In particular, we do not want to allow that a world where P is false is evaluation

[22] One might worry about (EW.1), for it implies that we can't give a good DBA showing that necessary truths or tautologies must receive credence 1. For let's say that I give $(P \vee \neg P)$ credence 0.9, rather than 1. Thus, I don't know that $(P \vee \neg P)$. Thus, (EW.1) says that we must allow evaluation worlds where $(P \vee \neg P)$ is false. But then we won't be able to show that there is a guaranteed loss from having such a credence. Note that since the evaluation worlds are all possible worlds, this worry does not arise. There are no evaluation worlds where $(P \vee \neg P)$ is false.

accessible from a c-world where the agent knows that *P*. If we adopt the EW proposal we can control for the agent's ignorance, without overcompensating and being unable to criticize agents that are criticizable. This is the motivation for EW.2.

There is, however, a loose-end to tie up. After all, EW.2 says that if the agent has the information that *P* at *some* time at which a set of bets are offered, then *P* is true at all the evaluation worlds for the set of bets. But the argument I just gave appealed only to one bet at one time. So, it might seem, I haven't really motivated EW.2, but instead a different condition, perhaps like the following:

ALL If the agent knows that *P* (in *c*) at ALL times over which the bet strategy ranges, then no world *e* where *P* is false is an evaluation world (relative to *c*).

What we have, I think, are two clear cases, EW.1 and ALL. The first says that if for no time over which a set of bets ranges the agent knows that *P*, then in some evaluation worlds for that set *P* is true and in some ¬*P* is true. The second says that if for all the times over which a set of bets spans the agent knows that *P*, then *P* is true in all evaluation worlds for that set.

The difficult cases are the ones where at some time over which the set of bets range, the agent knows that *P* and at some time he does not. This is exactly the kind of case we face when it comes to the *COND* DBA. The agent initially does not know that *E*, and then later does. There are two things that one might say about such cases:

LACK If the agent lacks knowledge that *P* (in *c*) at some time over which the bet strategy ranges, then there is at least one evaluation world (relative to *c*) where *P* is true, and one where *P* is false.

EW.2: If the agent knows that *P* (in *c*) at some time over which the bet strategy ranges, then no world *e* where *P* is false is an evaluation world (relative to *c*).

Note that each of these conditions seem to go against the motivation provided for one of EW.1 or ALL. If we go with LACK then we aren't able to criticize agents when they know a proposition is true and yet still bet in such a way that gain is possible only if that proposition is false. If we go with EW.2, we seem to make the other error. Let's say that you make a bet at *t0* on *E*, and then at *t1* there is another bet which is made if you learn *E*. According to EW.2, this set of bets is evaluated at worlds all of which have *E* true. This might seem unfair with respect to the bet made at *t0*. After all, that bet was made when the agent didn't know that *E*.

What should we say about this? The solution is to opt for EW.2, but pay special attention to how EW is used to evalute DBAs. When we do this, we will see that we don't end up evaluating bets incorrectly. To illustrate this point, consider the *COND* DBA again. We have two parts to the total betting strategy: {C1, C2, C3} and {C1, C2, C3}. EW.2 says that the set {C1, C2, C3} is only evaluated at worlds where *E* is true. The worry is that this is unfair to bets C1 and C2, since at the time these bets are considered, the agent doesn't know that *E*. But this worry actually does not arise. Although {C1, C2, C3} are evaluated only at worlds where *E* is true, this does not mean that bets C1 and C2 are evaluated at only such worlds. For when we look at {C1, C2, C3}, we see that this set is only evaluated at worlds where *E* is false. So, when we look at the bigger picture, we see that bets C1 and C2 are evaluated at worlds where *E* is true and where it is false. It is only C3 that is evaluated solely at worlds where *E* is true. But

this is perfectly acceptable, given the motivation outlined above, since the agent knows that *E* at the time C3 is considered.

I conclude, then, that EW.2 is well-motivated, as is EW.1. The main reason for adopting EW is because it gives a precise way of making sure the agent's guaranteed losses are not due to ignorance, but it does this without overcompensating and letting the agent off too easy.

6.9 Objection: Trivial Bets

There is a particularly surprising feature about the EW proposal. The purpose of this section is to show that this surprising feature does not lead to any adverse consequences, so long as we make a slight and plausible modification.

Condition (i) of the EW proposal allows as acceptable for good DBAs bet strategies that depend on information the agent doesn't have. The only requirement is that the strategy partitions the set of worlds. So, for instance, we are allowed a strategy for offering bets where we offer a bet on $\neg P$ if P and on P if $\neg P$. More explicitly, consider the following two bets:

Table 14: Bets Q1 and Q2

Q1	
P	$1 - n$
$\neg P$	*-n*

Q2	
P	*n – 1*
$\neg P$	*n*

Now consider the following strategy:

Table 15: Trivial Bet Strategy 1

{*w*: $\neg P$ true at *w*}	→	{Q1, <u>Q2</u>}
{*w*: *P* true at *w*}	→	{<u>Q1</u>, Q2}

This strategy partitions all possible worlds. And it looks like it guarantees losses. After all, we have set things up so that whichever bet is offered is a loser for the agent accepting the bet. If P is true, then Q2 is offered and the payout is $n - 1$. If $\neg P$ is true, then Q1 is offered and the payout is $-n$. As long as $n < 1$, this amounts to a certain loss. So, my framework appears to say that there is nothing wrong with a DBA for an agent that is probabilistically coherent! This would not be a welcome result.[23]

Thankfully, my proposal says no such thing, since so long as $n < 1$ and the Q1 or Q2 bet is accepted, the evaluation worlds for {Q1, Q2}and {Q1, Q2}will include worlds where the truth-value of P varies. Thus, we don't get the result that this is an acceptable DBA.

However, one might think there is a quick way to reassert the problem. Assume now that the true one of either P or $\neg P$ will be learned at *t1*, and consider the following bet, offered at *t1*:

Table 16: Bet Q3

Q3	
P	0
$\neg P$	0

This bet has an expected payout of 0 and so is condoned. Now consider the following strategy:

Table 17: Trivial Bet Strategy 2

{w: $\neg P$ true at w}	→	{Q1, Q2, Q3}
{w: P true at w}	→	{Q1, Q2, Q3}

[23] It is actually quite important that my proposal doesn't have this consequence. For if it did, then though we could establish the Dutch Book Theorem: "If you violate *COND*, then your doxastic state condones bets that are guaranteed to lose money," we could also show: "If you do not violate *COND* (or any other norms), then your doxastic state condones bets that are guaranteed to lose money." That is, we could not establish the Weak Converse Dutch Book Theorem.

The worlds that condone {Q1, Q2, Q3}are all worlds where ¬*P* is true, and where the agent learns ¬*P* at *t1*. Thus, the evaluation worlds for {Q1, Q2, Q3} are all worlds where ¬*P* is true. For a similar reason, the evaluation worlds for {Q1, Q2, Q3} are all worlds where *P* is true. Since Q3 doesn't change the agent's losses or gains, this strategy is one where the agent is certain to lose money. So, my framework appears to say that this is a good DBA. Again, this would not be a welcome result.

There is, however, a response to this. Note that Q3 is a trivial bet: no matter what happens, no money changes hands. It is plausible that we should rule out such trivial bets in DBA strategies. If we rule them out, then we have no such problem.

But one might think that we can reassert the problem without such a trivial bet. Imagine the following two bets offered at *t1*:

Table 18: Bets Q4 and Q5

Q4	
P	0
¬*P*	*n*/2

Q5	
P	–(*n* – *1*)/2
¬*P*	0

Now consider the following strategy:

Table 19: Trivial Bet Strategy 3

{*w*: ¬*P* true at *w*}	→	{Q1, Q2, Q4, Q5}
{*w*: *P* true at *w*}	→	{Q1, Q2, Q4, Q5}

Now, if ¬*P* is true, then ¬*P* is learned at *t1*, and so in all evaluation worlds ¬*P* is true. Thus, the payout from Q1 is –*n*, and the payout from Q4 is *n*/2. Putting these together, the total payout is –*n*/2, which is negative. If *P* is true, then *P* is learned, and so in all evaluation worlds *P* is true. Thus, the payout from Q2 is *n* – 1, and the payout from Q5 is –(*n* – *1*)/2. Putting these together, the total payout is (*n* – 1)/2, which is again negative. So, we appear to have guaranteed betting losses. Note further that Q4 and Q5

are not trivial in the sense that Q3 is trivial. It is possible that money changes hands according to these bets.

So, does my proposal say that if you are probabilistically coherent, then you are guaranteed to lose money? There are two reasons to think that the answer is *no*. First, note that Q4 and Q5 are not condoned, according to the way in which that term was defined. Neither Q4 nor Q5 have zero expected payout, as judged by the credence function to which those bets are offered. Further, there is no way to change Q4 and Q5 so that they have zero expected payout and are such that it is possible that money changes hands. So, there is no way to make Q4 and Q5 non-trivial and still have zero expected payout. So, there is no way to make Q4 and Q5 non-trivial and still be condoned by the agent's doxastic state.

One might claim, however, that my definition of 'condones' is too narrow. We should modify the definition to have the consequence that a credence function condones any bet that has expected payout *greater than or equal* to zero. Q4 and Q5 are both condoned in *this* sense, and perhaps that is all that matters. However, I think we lose something when we modify the definition in this way. For what we really want to see is which bets your credence function views as *fair*. Bets that have expected payouts not equal to zero aren't viewed as fair by your doxastic state, they are viewed as skewed to one side or the other. If we need to focus on bets that your credence function says are fair, then we should understand 'condone' as I originally understood it.

One might not moved be moved by this response. There is, however, a second one to be made. This response is to note that Q4 and Q5, though not trivial bets in the sense of Q3, are still trivial in an extended sense. Say that a bet is trivial relative to the

betting strategy if it is a bet such that if you eliminate that bet from all the strategies, but keep the evaluation worlds fixed as if that bet were still operative, then any part of the strategy that resulted in a gain still results in a gain, and any part of the strategy that resulted in a loss still results in a loss. In this sense, Q3, Q4, and Q5 are all trivial bets relative to their strategies. In what sense are they trivial? They are trivial in that they are added simply to gerrymander the evaluation worlds. The solution, then, is to rule out such trivial bets. We restate the general proposal by saying that goods DBA have strategies including no trivial bets, in this extended sense of 'trivial bet'. In the *COND* DBA, C3 is not a trivial bet in this sense, so the *COND* DBA still satisfies the EW proposal.

This response is consistent with the idea that we want to control for the agent's ignorance in Dutch Book situations. If a DBA has trivial bets in the extended sense (in the sense in which Q4 and Q5 are trivial), then the guaranteed loss of money is *solely* a result of the non-trivial bets. But then it is plausible to think that any error in the doxastic state must depend solely on the part of the doxastic state that condones those bets. But if that's where the error lies—in the part of the doxastic state that condones the non-trivial bets—then it is inappropriate to consider later bets made at later times. For doing so is simply a way to use information that the agent doesn't yet have against a set of credences that are supposed to be flawed *without* having that information. By not allowing trivial bets in this sense, we prohibit the agent's ignorance from being used against him.[24]

[24] One can see that the ignorance of the agent is being exploited in these examples more directly. The decision about whether to offer the agent Q1 or Q2 depends on the truth of *P* and yet has to be made by the bookie before the agent learns *P*. This is usually controlled for by allowing evaluation worlds where the truth of *P* varies. However, in this case the trivial bets prohibit this from happening.

We can see that this is the case by considering the possible consequences of removing trivial bets (as defined here). If we remove trivial bets this either will or will not alter the payouts of the bets. Suppose removal of the trivial bets does not alter the payouts of the bets. Then there is no harm in removing them, since the Dutch Book verdict is the same. Suppose, then, that removal does alter the payouts of the bets. Then, given the definition of trivial bets, this is because the evaluation worlds change. The only way in which the evaluation worlds change by *removing* bets is to increase the set of evaluation worlds. So the only purpose of the trivial bets was to restrict the set of evaluation worlds. But the set of evaluation worlds is restricted only by the agent gaining more information. So the only purpose served by the trivial bets is to ensure that the agent gets more information. Thus, the agent is guaranteed to lose solely because of bets made earlier, restricted to information he will get later. This is *not* the case with the DBA for PROB or *COND*. By ruling out trivial bets, we keep this sort of thing from happening.

6.10 Objection: The Role of Assumptions

In the introduction to this chapter, I noted that the proposal I would develop allows us to treat the *COND* DBA differently from the *Reflection*/*Self-Respect* DBAs. However, I also noted that after presenting the proposal, one is in a better position to consider the role that assumptions play in various Dutch Book Arguments. In this section I will consider this issue, and note how one can lodge an objection against my proposal by considering the assumptions in the arguments.

In the *COND* DBA, we made several assumptions, LA and PA. By making these assumptions, we essentially ignored certain kinds of worlds for the purposes of the

COND DBA. LA tells us to ignore worlds where E_i is true and the agent doesn't learn E_i. PA tells us that we're only considering worlds in which the things the agent could learn form a partition. We are to ignore any worlds where evidence is not like this.

According to the way I've advocated understanding DBAs, we start out with a set of possible worlds that must be partitioned by the DBA betting strategy. It is important to notice that this set of worlds is restricted in certain ways. Most obviously, this set of worlds is restricted to include only those worlds where the agent violates the norm in question. Assumptions like LA and PA further restrict the set of worlds that must be partitioned by the betting strategy. They tell us that we do not consider worlds where LA or PA are false. This might make one worry that by making these assumptions, we are stacking the deck against the agent, by unfairly restricting our consideration to a certain subset of all the worlds.

6.10.1 The Problem of Assumptions

This worry is made more pressing by noting that with the proper assumptions, we can get a good DBA for *Reflection*. After showing this, I will explain how to interpret this result within the framework I've constructed for evaluating DBAs.

Consider, again, an agent with *Reflection*-violating credences:

$\text{cr}_0(A|\langle \text{cr}_1(A) = r\rangle) = n$

$\text{cr}_0(\langle \text{cr}_1(A) = r\rangle) = d$

I explained above how my proposal shows that the *Reflection* DBA is not convincing. However, imagine that we alter the scenario by making the following assumptions:

RA1: If $\text{cr}_1(A) = r$, then the agent will learn at *t1* $\langle \text{cr}_1(A) = r\rangle$ as well as some P such that $\text{cr}_0(A|\langle \text{cr}_1(A) = r\rangle \wedge P) = r$,

RA2: If $cr_1(A) \neq r$, then the agent will learn at *t1* that $\langle cr_1(A) \neq r \rangle$.

Just as in the *COND* DBA, these assumptions rule out from the initial set of worlds any worlds that don't satisfy RA1 and RA2.

Why do I choose these two assumptions? According to RA1 and RA2, the agent is fully confident at *t1* of whatever the truth is concerning his credence at *t1* in *A*. Thus, the proposal I have sketched will give the result that these facts are held fixed in all evaluation worlds. Specifically, if $cr_1(A) = r$, then the offered bets are {R1, R2, R3}. Given RA1, $cr_1(\langle cr_1(A) = r \rangle) = 1$ and so every evaluation world for this set of bets is one where $cr_1(A) = r$. Thus, this set of bets guarantees a loss. If $cr_1(A) \neq r$, then the offered bets are {R1, R2, R3}. Given RA2, $cr_1(\langle cr_1(A) \neq r \rangle) = 1$ and so every evaluation world for this set of bets is one where $cr_1(A) \neq r$. Thus, this set of bets guarantees a loss, too.[25] Thus, the strategy for offering bets guarantees that the agent will lose money. So, we appear to have the makings of a good DBA for *Reflection*. What should be said in response to this?

6.10.2 Response to the Problem

My response is that something irrational is happening in this situation, although the blame cannot be pinned solely (or even primarily) on the violation of *Reflection*. To understand this response, it is useful to think of Dutch Book Theorem—which is the key premise in a DBA—as similar to a *reductio* argument. If we simplify it, the Dutch Book Theorem says:

[25] RA1 is more complex than RA2 because it ensures that the agent does not violate *COND* in getting the information about his own belief state when $cr_1(A) = r$. Were the agent to become certain of only $cr_1(A) = r$ at t_1, then the agent would be forced to violate *COND*. If this were the case, the guaranteed losses could

"If you violate principle *p*, then there are guaranteed betting losses."

Similarly, before drawing the final conclusion, a *reductio* argument can be presented as a conditional:

"If *reductio* hypothesis, then contradiction."

Of course, many *reductios* depend on auxiliary premises or assumptions in deriving the contradiction. So, more accurately, the *reductio* conditional looks like this:

"If *reductio* hypothesis and assumption 1, 2, …, *n*, then contradiction."

This conditional gives us good reason to reject the *reductio* hypothesis in proportion to the plausibility of the assumptions used. Similarly, the key premise of the DBA really should look like this:

"If you violate principle *p* and assumption 1, 2, …, *n* hold, then there are guaranteed losses."

My claim is that this gives us good reason to think that violation of *p* is irrational in proportion to the plausibility of the assumptions used. The best case is when we have *no* assumptions. The original DBAs for *Reflection* and *Self-Respect* appeared to be such cases. However, my proposal shows that this is mistaken: without assumptions the *Reflection*/Self-*Respect* DBAs fail. In this modified *Reflection* DBA we do have guaranteed losses, but I claim that the guaranteed losses tell more against the assumptions than against the violation of *Reflection*.

There are two reasons that we should be suspicious of the assumptions in this modified *Reflection* DBA and thus think that the guaranteed losses in this modified *Reflection* DBA do not show that violation of *Reflection* is irrational. First, the

---

be pinned on violation of *COND*, rather than *Reflection*. The extra complexity of RA1 allows us to avoid this.

assumptions are intuitively odd assumptions. In the *COND* DBA we need assumptions, because *COND* concerns evidence acquisition, and we must know what this acquisition is like. But in the *Reflection* DBA, this is not the case. *Reflection* doesn't concern evidence acquisition, so it is not clear why we would need assumptions that concerns this. We can bring out the oddity of the assumptions more clearly by noting that *Reflection* is a *synchronic* principle. It tells you something about what your conditional credences should be like *at one time*. Now, it is true that the credences in question are conditioned on propositions about other times, but the norm simply governs your conditional credences at one time. Despite its synchronic nature, the assumptions needed for the modified *Reflection* DBA are *diachronic* assumptions in that they concern some time *later* than the violation of *Reflection*. This should make us suspicious of the assumptions. In the *COND* DBA things are not like this. *COND* is a genuinely *diachronic* principle in that it tells you something about the relationship between credences at one time and credences at another. Thus, it is not surprising that in such a DBA we would need *diachronic* assumptions.

The second reason that one should be suspicious of the assumptions in the modified *Reflection* DBA is the kind of epistemic behavior to which they commit the agent. The assumptions dictate that the agent's beliefs change in a particular and odd way. For, given the assumptions, the situation is one where there is some P such that you *must* become fully confident in P whenever you become fully confident in $\langle \mathrm{cr}_1(A) = r \rangle$.

Consider first the simple situation where there is just one such P. First, if there is no necessary connection between this P and $\langle \mathrm{cr}_1(A) = r \rangle$, then it is pathological for an agent to be forced to have full credence in P whenever he has full credence in $\langle \mathrm{cr}_1(A) = r \rangle$. Suppose, then, that there is a necessary connection between the truth of P and $\langle \mathrm{cr}_1(A) = r \rangle$ so that the following is true:

(1) $\Box [P \leftrightarrow \langle \mathrm{cr}_1(A) = r \rangle]$

Now, (1) is either *a priori* or it is not. If it is *a priori*, then the agent must give it full credence. Accordingly, $\mathrm{cr}_0(P \leftrightarrow \langle \mathrm{cr}_1(A) = r \rangle) = 1$. Despite this, the situation is set up such that:

$\mathrm{cr}_0(A|\langle \mathrm{cr}_1(A) = r \rangle) \neq \mathrm{cr}_0(A|\langle \mathrm{cr}_1(A) = r \rangle \wedge P)$

But this is incoherent. So, if (1) is *a priori*, then the agent is incoherent, quite apart from his violation of *Reflection*. Consider the case where (1) is not *a priori*. In this case, (1) is an *a posteriori* necessary truth, perhaps like: $\Box [\mathrm{Water}(x) \leftrightarrow \mathrm{H_2O}(x)]$. But note that in this case, (1) is a very odd sort of *a posteriori* necessary truth. For it is *a posteriori*, and yet whenever the agent learns the truth of one side of the bi-conditional, he learns the other side. This makes it unlike any other example of an *a posteriori* necessary truth.

In summary, (1) is either true or not. If it is not, then RA1 encodes pathological epistemic behavior, for there is no necessary connection between P and $\langle \mathrm{cr}_1(A) = r \rangle$, and yet the agent acts as though there is. If (1) is true, then it is either *a priori* or *a posteriori*. Either way, we face a puzzle. So, I conclude that RA.1 itself encodes

inappropriate epistemic behavior. The betting losses in the modified *Reflection* DBA are not attributable to the violation of *Reflection*.[26]

If this is correct, then it helps to illuminate the role of assumptions in DBAs. The fact that a credence function condones bets that guarantee losses points to the fact that there is something irrational going on. But the blame may not be due to the violation of the norm in question. Given this, a possible response to the *COND* DBA is to claim that the guaranteed loss is not to be blamed on the violation of *COND*, but instead on the assumptions used to generate this guaranteed loss. Though this is a possible response, I think it is implausible when it comes to the *COND* DBA. It is true that the assumptions in the *COND* DBA do not perfectly describe every evidential situation, but I believe that they accurately describe certain sorts of evidential situations in a plausible way. The guaranteed losses in such situations thus reflect strongly back on the violations of *COND*. In the modified DBA for *Reflection*, this is not the case. As I have shown, the assumptions themselves are suspect.

6.10.3 Assumptions, Again

The response I have given above shows that if RA1 and RA2 are assumed to always hold—that is, if they are taken to be necessary—then there really is something irrational about the situation. But even if this is right, there is a different way of pressing the problem of assumptions. Instead of offering a Dutch Book for *Reflection*, while

[26] Note that I have shown this after making the simplifying assumption that there is one particular P that the agent learns whenever he learns $\langle \mathrm{cr}_1(A) = r \rangle$. But this is an overly strict way of understanding RA1. To get the modified *Reflection* DBA, all we need is the weaker claim that there is *some* P that is learned whenever the agent learns $\langle \mathrm{cr}_1(A) = r \rangle$. In Appendix B, I show that a similar, though more complex, line of argument shows that this assumption is to blame, too.

assuming RA1 and RA2 are necessary, we could have offered a Dutch Book for the following *conditional* principle, with *no* assumptions:

(RA1 ∧ RA2) → *Reflection*

This principle says that if RA1 and RA2 happen to hold, the agent must satisfy *Reflection*. Since on the assumption that RA1 and RA2 are necessary, we can show that any violation of *Reflection* leads to guaranteed losses, it follows that any violation of the principle just stated will also lead to guaranteed losses. But this is odd. The principle above doesn't seem like a principle of rationality. Further, nothing in the proposal that I have given allows us to explain this away, since there are no assumptions on which to lay the blame.

Something needs to be said about this problem if we are to have a full picture of how to understand Dutch Book Arguments. But note that this problem is a problem that arises for *any* understanding of Dutch Book Arguments, and not only for the framework I have proposed here. The root of the problem is that if you can give a Dutch Book for some principle, after making certain assumptions, then you can also give a Dutch Book for a conditional principle, with no assumptions of the form:

assumptions → principle

In no way does this issue arise for DBAs specifically in light of the framework that I have set up. So, for the purposes of this chapter, my response to this issue is the following: Either there is a solution to this problem for DBAs or there isn't. If there isn't, then *every* way of understanding DBAs is flawed. If there is, then I can adopt that solution and add it to the framework that I have presented here. So, at least for the purposes of this paper, I can sidestep this objection, for it is an objection to Dutch

Books *in general*, and not to my proposal about how to understand the particular issue that the EW proposal addresses.

6.11 Conclusion

In summary, here is the proposal that I offer as a necessary condition on a DBA being good:

> We have a good DBA *only if*
>
> (i) there is a strategy for offering bets, which includes no trivial bets, where each world is assigned to a set of bets, and this assignment partitions the space of worlds, and
>
> (ii) the agent's credence function condones each of those bets at the appropriate time, and
>
> (iii) the set of bets leave the agent with a loss in all the evaluation worlds for that set of bets as determined by EW.

> **EW:** A world *e* where ¬*P* is true is evaluation inaccessible from c-world, *c*, *iff* the agent knows that *P* (in *c*) at some time over which the bet strategy ranges

This proposal builds on the Christensen/Briggs idea that all guaranteed betting losses are not to be treated the same. Some guaranteed betting losses matter and some do not when it comes to Dutch Books. I have implemented that idea by claiming that there are worlds evaluation accessible from condoning worlds that are not themselves condoning worlds. EW then builds off the familiar idea that DBAs are not convincing if they exploit the agent's ignorance. Putting these together we have a proposal that is

intuitively satisfying. According to this proposal we can rightly criticize the DBAs for *Reflection* and *Self-Respect*. However, the DBA for *COND*, as well as the DBA for *Finite Additivity* come out unscathed.

CHAPTER 7

EVIDENCE AND DISSOCIATION

7.1 Introduction

In Chapter 6 I defended a way of understanding Dutch Book Arguments according to which the DBA for *Reflection* and *Self-Respect* can be rejected without rejecting the DBA for *COND*. In this chapter, I would like to step back from the details of that proposal and look at the general structure of the response to the *Reflection*/*Self-Respect* DBAs.

The essence of the response to the *Reflection*/*Self-Respect* DBAs, which is shared by Briggs, Christensen, and my EW proposal, is that an agent is not to be criticized for having inaccurate beliefs about his own doxastic states. This point is especially clear in Briggs's ([2009]) presentation of her proposal. Recall that according to Briggs we are to abstract away from the identity of the believer when we evaluate the rationality of the believer's doxastic states. We do this by allowing propositions about the agent's doxastic state in condoning worlds to vary in truth-value at evaluation worlds. Christensen strikes a similar chord. And though the proposal I give for evaluating DBAs is different than either of these proposals, the same idea is present: a believer need not be criticizable for failing to have information about her own doxastic state.[1]

[1] According to the EW proposal, the failure of the *Reflection*/*Self-Respect* DBA depends on the possibility that the agent fails to have information about her own doxastic state. Since we don't want the *Reflection*/*Self-Respect* DBA to fail because of some *other* irrationality on the part of the agent, the agent's failure to have this information must not be irrational.

Now, it often seems that there *is* nothing irrational about failing to have information about one's own doxastic states. At other times, however, there is something odd about this. For example, consider Tom, who believes the propositions:

(1) *A*.

(2) It's not the case that Tom believes *A*.

If we abstract away from who is doing the believing here—if we ignore the fact that it is Tom who believes (1) and (2)—there is nothing problematic with believing (1) and (2). By doing this, however, we lose something. For it seems important to our example that it is *Tom* who believes (1) and (2). At least sometimes, this fact seems to matter in the evaluation of the rationality of Tom holding those beliefs.

Now, the response to the *Reflection*/*Self-Respect* DBA doesn't commit one to the claim that one can *always* be rationally ignorant of one's own doxastic states. But it does commit one to saying that one can *sometimes* be rationally ignorant of one's own doxastic states. Thus, the kind of response offered to the *Reflection*/*Self-Respect* DBA in chapter 6 brings up a natural question: when is it rationally permissible to be ignorant of one's own doxastic state? This question is important because the more that an agent is ignorant of his own doxastic state, the less the agent seems to exhibit a certain kind of desirable unity to his representation of the world. The agent's first-order doxastic states are a certain way, and yet the agent's second-order doxastic states think that they are a different way.

I think the most plausible answer to this question about when it is rationally permissible to be ignorant of one's own doxastic state is the following:

The Middle-of-the-Road Answer: Sometimes it is rationally permissible for an agent to be dissociated from her own doxastic states, and sometimes it is not.[2]

Though plausible, this answer is also disappointingly vague. In what follows, I will show how an account of evidence (perhaps like RAE*) can be of considerable help here. Appealing to such an account allows us to give a more precise answer to when an agent can be rationally dissociated from her own doxastic states. However, there are further benefits to saying something about this question. In doing so, we will be able to explain why Moore's Paradox-like doxastic states strike us as irrational. Further, we will be able to explain when something like *Self-Respect* is rationally binding on an agent. Finally, we will be able to explain the allure of the DBA for *Self-Respect* despite the fact that it is flawed.

7.2 In Favor of Limited Dissociation

Before showing these things, it is important to say something about why The Middle-of-the-Road Answer seems appropriate. The response to the *Reflection*/*Self-Respect* DBA allows that higher-order credences "float free" from the lower-order credences: just because $\mathrm{cr}(A) = n$, nothing follows about the value of $\mathrm{cr}(\langle \mathrm{cr}(A) = n \rangle)$. If we say nothing further, then we implicitly say that it is *always* rationally permissible for an agent's higher-order credences to be dissociated in this way from the lower-order ones. This is to reject The Middle-of-the-Road Answer.[3] There is, I think, good reason not to do this.

[2] Note that this is all that is needed in the response to the *Reflection*/*Self-Respect* DBAs. As long as it is sometimes permissible, violation of *Self-Respect* or *Reflection* does not *guarantee* betting losses.
[3] Which is not to say that either Christensen or Briggs think it is always rationally permissible for an agent's higher-order credences to be dissociated from the lower-order ones.

Recently, there has been an awareness of the importance of so-called higher-order evidence. The kind of higher-order evidence I'd like to focus on is the kind of evidence that leads a rational agent to doubt the accuracy of a first-order doxastic state. The growing literature on peer disagreement is a striking example of this. In a typical case where I learn that a peer disagrees with me about an issue even though we possess the same body of evidence, this gives me information about that body of evidence. That is, the disagreement provides me with a reason to think that my initial evidence isn't quite as compelling as I thought. According to many, this should result in me giving less credence to the target proposition I formerly believed to be supported by the evidence. Another example of higher-order evidence, not involving peer disagreement, is provided by Arntzenius's ([2003]) Shangri-La case. In that case, the agent gets evidence that his memory has been tampered with, which gives the agent some reason to doubt that his credences concerning past events are accurate. Christensen ([2007]) provides yet a further case. In this case the agent is given evidence that she's been slipped a logic-distorting drug. This, in turn, leads the agent to doubt that her own credences about logical truths are accurate. And in Christensen ([*forthcoming*]) there are further examples of higher-order evidence. One is a case where a doctor forms some medical decisions and then is informed that he has been awake for 36 hours. This evidence about how long he has been awake provides the doctor with some reason to doubt that his own medical opinions are accurate. In general, then, higher-order evidence is evidence that gives one reason to doubt one's cognitive functioning, leading to rational doubts about the cogency of some of one's first-order doxastic states.[4]

[4] Some of the examples of undercutting defeaters considered in Chapter 5 are naturally thought of as examples of higher-order evidence, too.

In the examples above, the stories are told so as to make it seem rational for the agent to adjust his first-order credence upon learning something about how he acquired that first-order credence. That is, upon having some high (second-order) credence that something has gone wrong with a first-order credence, we are urged to conclude that this should have an affect on the first-order credence.

Let 'B[x]' be a predicate that takes a proposition about a doxastic state as argument, and says of it that there is something epistemically inappropriate about it. In the cases of higher-order evidence considered we have agents that receive evidence which ultimately leads to a credence of the form:

(HOC) $\mathrm{cr}(\mathrm{B}[\langle \mathrm{cr}(A) = n \rangle]) = \text{high}$.

For instance, as a result of learning that one has been slipped a logic-distorting drug, one might be such that $\mathrm{cr}(\mathrm{B}[\langle \mathrm{cr}(P \rightarrow (Q \rightarrow P)) = 1 \rangle] = \text{high}$. In such a case, the higher-order evidence is plausibly taken to be the proposition that you have been slipped a logic-distorting drug (or the proposition that you have been told you've been slipped a logic-distorting drug). The credence that meets HOC schema is not naturally described as the higher-order *evidence*, but rather one of the *effects* of such evidence. Now, one might hold a view according to which possession of higher-order evidence need not lead to something like HOC. Further, one might hold that having a credence like HOC need not lead one to change one's first-order credences. But it is plausible to maintain both of these. That is, upon receiving the higher-order evidence, one rationally acquires a credence like that in HOC, and such a credence should have an effect on the first-

order credence mentioned in HOC.[5] It seems that Arntzenius and Christensen think that acquiring something like HOC should have an affect on one's first-order credences. Christensen ([2007]), for instance, thinks that this gives us reason to think that rational agents need not give credence 1 to tautologies.

It is plausible, then, to hold that higher-order credences that meet the HOC form should influence lower-order credences. But if one accepts that HOC-like credences should have this effect, while also holding that an agent's higher-order credences can be completely dissociated from his lower-order credences, then there would be an odd asymmetry. According to such a view, second-order credences make rational demands on first-order credences. So, if I believe there is something inappropriate about my first order credences, this rationally motivates a change in my first-order credence. However, first-order credences do not make rational demands on second-order credences. What is odd about such an asymmetry? Well, such an asymmetry implies that there is a close rational connection between one's second- and first-order credence going in one direction, but not in the other direction. In virtue of having some second-order credences about the quality of one's first-order credences, one's first-order credences must change. That is, one's first-order credences must be responsive to one's second-order credences. But despite this, one's second-order credences can be rationally unresponsive to one's first-order credences. It is my contention that a more accurate picture of rational agents will reveal more symmetry than this. This gives some reason to prefer The Middle-of-the-Road Answer.

[5] Of course, some disagree with this. For instance, Thomas Kelly ([2005]) defends a view according to which this is not the case.

7.3 Evidence and Limited Dissociation

So, what kind of constraints might lower-order credences place on higher-order ones? An account of evidence, can provide a promising start toward answer this question. The simple idea I'd like to propose, is that it is rationally inappropriate for an agent to be dissociated from elements of her own doxastic state when those elements are themselves evidence for her. This provides one way in which first-order credences could place constraints on second-order ones. In particular, an account of having evidence like RAE* will tell us that if one has a reliable route to one's first-order credences, then that one has such credences is evidence for one, and rationality demands that one respond to this evidence by harboring responsive second-order credences about such first-order credences. This would forge a rational connection going from first-order credences to second-order credences.

Note that this is true whether or not we go with RAE* or RAE*-t. If we go with RAE*-t, then all evidence must be true; if we go with RAE* then this is not the case. Nevertheless, whenever there is a reliable indication of a first-order credence, both RAE* and RAE*-t say that that the agent has such a first-order credence is evidence for the agent. Accordingly, the agent's second-order credences are required to be responsive to facts about the agent's first-order credences.

It is important to point out that, on this kind of account, the demand that first-order credences make on second-order credences is a merely contingent demand. It is not the case that all of an agent's first-order credences in all situations place such a demand on the agent. Propositions about the agent's first-order credences must be *evidence* for the agent before this happens. But, on reflection, this seems just right. It is

not *always* the case that one's higher-order credences must be responsive to one's lower-order ones. This is why the Middle-of-the-Road Answer is the plausible one.

However, one might think that this displays an asymmetry between the relations that obtain between first-order and second-order credences: whereas the demand that first-order credences make on second-order credences is contingent, one might think that the demand that second-order credences make on first-order credences is not contingent. That is, one might think that whenever one has a credence like HOC, one is required to alter one's first-order credence. One could hold such a view. However, I have defended an account according to which this is false: second-order credences about the quality of one's first-order credences do not always require one to change one's first-order credences. So I hold that the rational demands that first-order credences place on second-order credences and that second-order credences place on first-order credences are all contingent.

In chapter 5, I discussed a very particular sort of situation where one has a credence like HOC. These were situations where one came to have a high credence in a proposition about how one came to have full credence in one's evidence. To illustrate, let *E* be some evidence that I have and let *UE* = "The way in which I came to believe *E* was unreliable." Having a high credence in *UE* is to have a credence meeting the HOC schema. I argued in Chapter 5 that for most of us, coming to have a high credence in *UE* will result in *E* being expunged from our evidence sets. So, high credence in *UE* (a second-order credence) rationally demands that one change one's first order credence (specifically, a high value for cr(*UE*) requires *E* to be expunged from the evidence set and so to go from receiving credence 1 to something less than 1). However, within that

restricted context, I argued that it is still reliability considerations that allow higher-order credences to have the effect that they do. There are bound to be agents and situations for which this does not hold. Imagine an agent who pathologically comes to have HOC-style credences all the time. For such an epistemic hypochondriac, I argued, a high value for cr(*UE*) does not necessarily expunge *E* from the evidence set. If that's right, then second-order credences like HOC do not always demand changes in first-order credence. So, at least in that restricted context, I've given a view according to which the normative requirements that second-order doxastic states make on first-order ones is itself a contingent matter. First-order doxastic states and second-order doxastic states contingently make rational demands on each other. Usually, for agents like us, in situations like we face, high credence that one's first-order credence is mistaken is a good reason to adjust that credence. Similarly: usually, for agents like us, in situations like we face, first-order credences demand that our second-order credences are responsive to them in some way.

7.4 Applications of RAE

I have explained how an account of having evidence will allow that lower-order credences sometimes place constraints on higher-order credences. Further, an account of evidence can tell us when such credences do this. In this final section, I'll demonstrate how this can shed light on several issues mentioned at the introduction to this chapter.

7.4.1 Moore's Paradox

Consider a probabilistic Moore's Paradox case. An agent might have the following credences:

$$\text{cr}(A) = 0.9 \qquad \text{cr}(\langle \text{cr}(A) = 0.9 \rangle) = 0.1$$

Such a distribution of credences is mildly odd. The case strikes one as significantly more paradoxical when we add to the case that the agent truly asserts, "*A*, but I don't believe *A*," which would seem to be permitted given those credences.[6] Why does the sense of paradox increase when we add this? An account of evidence like RAE* can be of help here. According to that account, the distribution of credences above is unacceptable if $\langle \text{cr}(A) = 0.9 \rangle$ is evidence for the agent in question.[7] Since we're assuming $\text{cr}(A) = 0.9$, both RAE* and RAE*-t say that $\langle \text{cr}(A) = 0.9 \rangle$ is evidence just in case the agent has a reliable route to its truth. I conjecture that when we initially examine that distribution of credences, we tacitly assume that the agent is somehow isolated from his first-order credences. If so, then RAE* says that $\langle \text{cr}(A) = 0.9 \rangle$ is not evidence and so that distribution of credences is not judged irrational. However, once we add to the story the agent's assertion, it becomes very difficult to imagine how the agent doesn't have a reliable route to the fact that $\text{cr}(A) = 0.9$. After all, a true assertion of that sentence seems to come from the fact that $\text{cr}(A) = 0.9$ and that $\text{cr}(\langle \text{cr}(A) = 0.9 \rangle) = 0.1$. This suggests that the agent isn't isolated from these features of his doxastic state, which suggests that the agent *does* have a reliable route to the fact that $\text{cr}(A) = 0.9$.

[6] I assume here that you are permitted to sincerely assert a proposition when you have a suitably high credence in that proposition.

Accordingly, RAE* says that this is evidence for the agent. Though there is no inconsistency or incoherence in that distribution of credences, there is still something epistemically inappropriate about it. RAE* provides an explanation of this.[8]

7.4.2 Self-Respect

Consider, now, the principle that Christensen calls *Self-Respect*:

Self-Respect **:** $\mathrm{cr}_t(\mathrm{A}|\langle \mathrm{cr}_t(\mathrm{A}) = n\rangle) = n$

Christensen ([2007]) has noted that satisfaction of *Self-Respect* is entailed by the satisfaction of PROB if the agent is fully confident and accurate with respect to his own credences. So, for instance, if $\mathrm{cr}_t(P) = n$ and $\mathrm{cr}_t(\langle \mathrm{cr}_t(P) = n\rangle) = 1$, then it must be that $\mathrm{cr}_t(P|\langle \mathrm{cr}_t(P) = n\rangle) = n$, which is *Self-Respect*. So, if you happen to be accurate and confident in your own credences, then you must satisfy *Self-Respect*, independent of any Dutch Book considerations. This seems to show that *Self-Respect* is a normative requirement, albeit in a restricted set of situations. But, as Christensen notes, this satisfaction of *Self-Respect* is merely accidental satisfaction. If I happen to be fully confident with respect to *your* credences, and I happen to think that your credences are identical to mine, then I must satisfy an interpersonal version of *Self-Respect*. That is, if $\mathrm{cr}_{\mathrm{me}}(P) = n$ and $\mathrm{cr}_{\mathrm{me}}(\langle \mathrm{cr}_{\mathrm{you}}(P) = n\rangle) = 1$, then it must be that $\mathrm{cr}_{\mathrm{me}}(P|\langle \mathrm{cr}_{\mathrm{you}}(P) = n\rangle) = n$. This interpersonal version of *Self-Respect* doesn't seem to tell us anything interesting about rationality. So, the thought goes, the regular version of *Self-Respect* doesn't tell us

---

[7] Note that we can get similar situations with logically weaker propositions, e.g., $\langle 0.85 < \mathrm{cr}(A) < 0.95\rangle$. One might think that these logically weaker propositions are more likely to be evidence than something like $\langle \mathrm{cr}(A) = 0.9\rangle$.

[8] I don't mean to imply that *only* RAE* can provide an account of this. The main point is that some account of evidence, will be helpful here. I focus on RAE*, because this is the account I have defended and investigated in detail. However, an account of evidence that makes some notion of availability crucial

anything interesting about the relation between first- and second-order credences, even in the restricted set of circumstances.

Is there anything more informative we can say about when one ought to satisfy *Self-Respect* that makes it look more interesting than the interpersonal version? By appealing to a normative account of what it is to have evidence, I think we can. RAE* is one such account. Though one need not adopt that particular account, I will appeal to some basic features of RAE* in what follows.

Consider RAE*-t. According to that account, a proposition is evidence for an agent if it is reliably indicated to the agent and is true. Thus, if $\langle \mathrm{cr}(P) = n \rangle$ is evidence for me, then $\mathrm{cr}(P) = n$ and $\mathrm{cr}(\langle \mathrm{cr}(P) = n \rangle) = 1$. Thus, if we endorse something like RAE*-t, an agent must satisfy *Self-Respect* with respect to some proposition, when his credence in that proposition is evidence for him. To point out that *Self-Respect* must be satisfied if an agent *happens* to be confident and accurate about his own credences makes satisfaction of *Self-Respect* a purely accidental feature of having probabilistically coherent beliefs. However, by having a normative theory of evidence standing behind it, we can say that there are situations where S *should* be confident and accurate about his own credences. In such situations, satisfaction of *Self-Respect* is required in a non-accidental way.[9] This helps to bring out a difference between satisfying *Self-Respect* for myself, and satisfying the interpersonal version. For if $\langle \mathrm{cr}_{\mathrm{you}}(P) = n \rangle$ is evidence for me,

---

to the having of evidence would be able to co-opt the explanation given in the text, including accounts that are more internalist.

[9] This point can be put in an alternative way if one is committed to hp-*COND* as an update rule. If an agent obeys hp-*COND*, then $\mathrm{cr}_t(\bullet) = \mathrm{hp}(\bullet|E_t)$, where E_t is the agent's evidence set at t. Thus, an agent is required to satisfy *Self-Respect* if (i) $\mathrm{hp}(P|E_t) = n$ and (ii) $\mathrm{hp}(\langle \mathrm{cr}_t(P) = n \rangle|E_t) = 1$. Of course, this could happen accidentally. However, if $\langle \mathrm{cr}_t(P) = n \rangle$ is evidence then (i) and (ii) are satisfied non-accidentally. For if $\langle \mathrm{cr}(P) = n \rangle \in E_t$, then (ii) is satisfied, and according to RAE*-t, if $\langle \mathrm{cr}_t(P) = n \rangle \in E_t$, then $\mathrm{cr}_t(P) = n$, which is equivalent to (i).

it doesn't follow that I must satisfy the interpersonal version of *Self-Respect*. This only follows if $cr_{me}(P) = n$, too. So we have here a real difference between the two principles. If your own credences are evidence for you, then you must satisfy *Self-Respect*. If your friend's credences are evidence for you, you need not.[10]

Now, if we opt for RAE*, rather than RAE*-t, we don't get quite these results. If we go with RAE*, then there will be situations where $\langle cr(P) = n\rangle$ is evidence for an agent, even though $cr(P) = m \neq n$. In such cases, the agent is not required to satisfy *Self-Respect*. In fact, in such situations, the agent is required to *not* satisfy *Self-Respect*, since if $cr(\langle cr(P) = n\rangle) = 1$ and $cr(P) = m$, probabilistic coherence mandates that $cr(P|\langle cr(P) = n\rangle) = m$. Nevertheless, since RAE* requires high reliability for evidence, these situations will be rare. It will almost always be the case that an agent must satisfy *Self-Respect* with respect to some proposition when his credence in that proposition is evidence for him. Thus, RAE* approximates the results that RAE*-t gives.

Christensen ([2007]) has also pointed out that there are situations where one must come close to satisfying *Self-Respect*, even if one isn't fully accurate or fully confident in one's own credences. In particular, if we let $cr(\langle cr(P) = y\rangle) = z$ and if $cr(P) = x$, then it must be that:

[10] In chapter 6 I argued that the general DBA for *Self-Respect* is not convincing. According to what we've just said, however, *Self-Respect* is a requirement with respect to any proposition P for which $\langle cr(P) = n\rangle$ is evidence for you. Thus, we would like to be able to get a DBA for *Self-Respect* under these conditions. Note that in such a situation, the agent will have the information about his doxastic state and so according to the EW proposal will be susceptible to the *Self-Respect* DBA. This is a nice feature of the proposal I made for evaluating DBAs, but note that it doesn't really restore a role for the *Self-Respect* DBA. For consider the class of situations where the *Self-Respect* DBA would be convincing. In all these situations the agent violates *Self-Respect*. Further, he either responds to his evidence or he does not. If he does not, then he is criticizable for this, and the *Self-Respect* DBA is not needed. If he *does* respond to his evidence (and continues to violate *Self-Respect*), then he is probabilistically incoherent, and so the *Self-Respect* DBA is not needed. So, the *Self-Respect* DBA ends up being convincing just when it is superfluous.

$$\frac{x + z - 1}{z} \leq \text{cr}(P|\langle \text{cr}(P) = y\rangle) \leq \frac{\min[x.\ z]}{z}$$ [11]

As one can see, as z approaches 1 the value of $\text{cr}(P|\langle \text{cr}(P) = y\rangle)$ converges on x. Since as one's accuracy improves, y approaches x, we get approximate satisfaction of *Self-Respect*. Now, one might wonder if an account of having evidence has anything to say about partial satisfaction of *Self-Respect*. In fact, we can capture the idea that one should come close to satisfying *Self-Respect* the more evidence one has about one's doxastic state.

Often, an agent will not have as evidence particular facts about her own credence function. Rather, more general facts about her own credence function will be evidence (see, for instance, footnote 7). This is especially clear if we adopt something like RAE*-t: though I may not have a reliable route to the fact that I have some precise credence in P, I may nevertheless have a reliable route to the fact that my credence in P is within some range. More specifically, it might be that though $\langle \text{cr}(P) = n\rangle$ is not evidence for me, $\langle \text{cr}(P) = n \pm \delta\rangle$ is evidence for me. If that's the case, then agents with such evidence must approximate *Self-Respect*. In particular, if $\langle \text{cr}(P) = n \pm \delta\rangle$ is evidence, then it should be that $\text{cr}(\langle \text{cr}(P) = n \pm \delta\rangle) = 1$. From this it follows that

[11] Christensen doesn't present his point in this way, but it follows from things that he says. Proof: $\text{cr}(\text{P}|\langle \text{cr}(P) = y\rangle) =_{\text{df.}} \text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)/\text{cr}(\langle \text{cr}(P) = y\rangle) = \text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)/z$. So, $\text{cr}(\text{P}|\langle \text{cr}(P) = y\rangle)$ is largest when $\text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)$ takes it maximum value, and is smallest when $\text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)$ takes its minimum value. The maximum value of $\text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)$ is the lower of the two values that cr(P) and $\text{cr}(\langle \text{cr}(P) = y\rangle)$ take since the probability of a conjunction must be less than or equal to the probability of each conjunct. This gives us an upper bound on $\text{cr}(\text{P}|\langle \text{cr}(P) = y\rangle)$ of $\min[\text{cr}(\text{P}), \text{cr}(\langle \text{cr}(P) = y\rangle)]/z = \min[x, z]/z$. The minimum value that a conjunction can take can be given as a function of the values of each conjunct as $\text{cr}(\text{A} \wedge \text{B}) \geq \text{cr}(\text{A}) + \text{cr}(\text{B}) - 1$. Intuitively, this is because the amount of $\text{cr}(\text{A}) + \text{cr}(\text{B}) - 1$ measures the amount that A and B must overlap. If $\text{cr}(\text{A}) + \text{cr}(\text{B}) \leq 1$, then they need not overlap at all and $\text{cr}(\text{A} \wedge \text{B})$ could be 0. But if $\text{cr}(\text{A}) + \text{cr}(\text{B}) > 1$, then there must be some overlap. The minimum amount of such overlap is given by $\text{cr}(\text{A}) + \text{cr}(\text{B}) - 1$. Given this, it follows that lower bound on $\text{cr}(\text{P} \wedge \langle \text{cr}(P) = y\rangle)$ is $(\text{cr}(\text{P}) + (\text{cr}\langle \text{cr}(P) = y\rangle) - 1)/z = (x + z - 1)/z$.

$\text{cr}(P|\langle \text{cr}(P) = n \pm \delta\rangle) = \text{cr}(P)$. If we're going with RAE*-t, since $\langle \text{cr}(P) = n \pm \delta\rangle$ is evidence, it follows that it is true that $\langle \text{cr}(P) = n \pm \delta\rangle$, and so we get that:

Self-Respect-Range[12]**:** $\text{cr}(P|\langle \text{cr}(P) = n \pm \delta\rangle) = n \pm \delta$

Self-Respect-Range is a generalization of *Self-Respect*. It tells us that if you have as evidence that your own credence in P is within the range of $n - \delta$ and $n + \delta$, then the value of $\text{cr}(P|\langle \text{cr}(P) = n \pm \delta\rangle)$ must be within the range of $n - \delta$ and $n + \delta$.

7.4.3 The Allure of DBAs

I will conclude by showing how an account of having evidence allows us to explain why the DBAs for *Reflection* and *Self-Respect* are so alluring, even though ultimately mistaken. Consider first *Self-Respect*. In the DBA for *Self-Respect*, Bet R3 is a bet on A at certain odds. But, I conjecture, it is natural to see the taking of a bet as an expression of a credence. For instance, imagine that you ask me if I think it is fair to take a bet on the Red Sox winning at 3:1 odds, and I say that I do, but that I'd go no higher. This is naturally seen as a way for me to express that $\text{cr}(\textit{Red Sox win}) = 0.75$. Now, imagine a different scenario, where you ask me what my credence is that the Red Sox will win, and I answer by telling you that my credence is 0.75 that they will win. In this second scenario, it is hard to imagine that I tell you this, and yet am not reliably related to this fact about my credence. There are undoubtedly situations where we misreport our credences. But there is at least a presumption in favor of thinking that me reporting my credence in a proposition is based on the fact that my credence in that proposition is what I report.

[12] As above, this is worked out under the assumption of RAE*-t. If we adopt RAE* then we get something close to this.

Now, consider what this tells us. First, the explicit taking of a bet is naturally understood as an expression of the fact that one has a particular credence. Second, the expression of a particular credence is often an indication that the agent has reliable access to the particular credence in question. So, if an agent takes a bet on A at odds n:1, where $r = n/(n+1)$, it is hard to imagine that the agent doesn't also have as evidence $\langle \text{cr}(A) = r \rangle$. And, as we saw above, if the agent has as evidence $\langle \text{cr}(A) = r \rangle$, then he must satisfy *Self-Respect*. What we have, then, is the betting scenario in the DBAs priming us to *not* consider the very kinds of situation where violation of *Self-Respect* is rationally acceptable: situations where the agent lacks reliable access to the values of his own credence function. This, I think, explains the allure of such a DBA.

However, there are good reasons not to understand a DBA as involving an agent actually buying or selling bets. First, the right way to present a DBA is by having various credences condone various bets. For instance, $\text{cr}(A) = r$ condones Bet R3 in the DBA for *Self-Respect*. But your credence value can condone a bet without you being willing to actually take the bet in question. You could, for instance, be averse to betting. The betting scenario, complete with a bookie and bettor, as Christensen ([1991]) has pointed out, is a bit of fiction not essential to the DBA. So, we can have bets that are condoned, even if the agent wouldn't actually take the bet if the bookie were to approach him. The same point, however, shows us that one can condone a bet even if one is very ignorant about the feature of one's belief state that leads to this condoning.

Second, there are situations where an agent might take a certain bet, even if his belief state did not condone it in the way needed for the DBA. One such situation is where the agent is simply being irrational: the agent is such that $\text{cr}(\textit{Heads}) = 0.5$, he is

certain of this (so $\text{cr}(\langle \text{cr}(\textit{Heads}) = 0.5 \rangle) = 1$) and yet pays \$1 for a bet on *Heads* that will pay him \$0.25 if *Heads* and nothing otherwise.

But there are other situations where an agent might take a certain bet though his belief state does not condone such a bet. Consider a scenario where you are uncertain about your credence in *P*, but you are forced to bet on *P*. In such a situation you might think that the best policy it to bet based on your expectation of the value that cr(*P*) takes:

$$\sum_i n_i \times \text{cr}(\langle \text{cr}(P) = n_i \rangle)$$

For instance, perhaps you aren't certain about the value of cr(*P*), in the sense that $\text{cr}(\langle \text{cr}(P) = 0.8 \rangle) = 0.5$ and $\text{cr}(\langle \text{cr}(P) = 0.4 \rangle) = 0.5$. In such a situation, the agent might legitimately take a bet on *P* that would be condoned by $\text{cr}(P) = 0.6$. This could be so even if the actual value of cr(*P*) is not 0.6.

So, reasonably taking a bet is different than having a doxastic state that condones the bet. We should thus clearly distinguish the two. Failing to do this in the case of the *Self-Respect* DBA, makes it more persuasive than it otherwise should be. By thinking of the DBA as a situation where the agent actually takes a bet, this leads us to think that the agent has elements of his doxastic state as evidence. And if this is the case, then the agent really does have a deficiency in his doxastic state. But one can be subject to a Dutch Book without having elements of one's doxastic state as evidence. In such situations, Dutch Book vulnerability is not necessarily a good guide to epistemic failure.

7.5 Conclusion

In this chapter I noted that the response to the *Reflection*/*Self-Respect* DBAs given in chapter 6 raises the question of what the appropriate relation is between first- and second-order doxastic states. I argued that a good picture of this relationship would show both kinds of doxastic states effecting each other and argued that an account of evidence can be of help in explaining this. RAE*, I think, performs nicely in these situations. It strikes the appropriate balance between requiring that agents have total access to their own doxastic state and allowing that agents can be rationally completely ignorant of their own doxastic state. I concluded by showing how appealing to an account of evidence like RAE* has other benefits: it helps to explain why probabilistic versions of Moore's Paradox can seem paradoxical; when *Self-Respect* is non-accidentally binding on an agent; and can help to explain why the *Self-Respect* DBA is so alluring, even if ultimately flawed.

CHAPTER 8

CONCLUSION

The primary goal of this dissertation has been to get clear on what it is to *have* evidence. That is, I have attempted to provide an answer to The Evidence Question:

> When, and under what conditions, does an agent have proposition, *E*, as evidence at *t*?

I have defended a particular answer to this question:

> **RAE*:** The set of propositions $\mathbf{E_t}$ is S's evidence set at *t* iff
>
> (i) for each $P_i \in \mathbf{E_t}$ there are reliable belief-forming processes, $\mathbf{p}_i$ available to S at *t* such that if S applied those operations S would believe all the P_i at *t* and those beliefs would be caused by reliable belief-forming processes,
>
> (ii) none of the $\mathbf{p_i}$ are inductive inference from S's other beliefs, and
>
> (iii) $\text{pr}(\neg(\text{b}P_i \wedge \mathbf{p_i})|\neg P_i) \geq \text{s}$.[1]

In the preceding chapters, many particular considerations in favor of RAE* and against RAE* have been considered. Instead of surveying these considerations, I would like to look at the bigger picture, and say something about why I think RAE* is a promising account of having evidence.

Consider two extreme views about evidence. According to one, your evidence includes all the true propositions. Such a view makes it clear how evidence connects an agent with the world, how evidence can justify true beliefs about the world, and why evidence is a good thing to update on. But such a view fails miserably in accounting for the fact that different agents have different evidence, and the fact that what your

evidence *is* depends, in some way, on what cognitive tools you are working with. It also fails to account for the fact that evidence is something that agents can use as a means to further truths. A view on the opposite extreme says that your evidence includes all the propositions about how things seem to you of which you are certain. This view accounts for the fact that different agents have different evidence. It also gives us a better picture of evidence as something that agents can use as a means to further truths: propositions about how things seem to you, the thought goes, are accessible and able to be used in a way that truths about the world are not. However, this view fails miserably in accounting for how evidence connects an agent with the world, how evidence can justify true beliefs about the world, and why evidence is a good thing to update on.

Throughout this dissertation, I have worked with the idea that an account of evidence should try to navigate between these two extremes. I think that RAE* strikes an appropriate balance. Reliability assures us that evidence connects an agent with the world, is usually true, and is a good thing to update on. The fact that reliability is referenced to the processes of belief-formation that the agent has ensures us that the evidence an agent has depends on the agent's position in the world and the agent's cognitive tools. This general feature of RAE* is one that makes for an attractive account of having evidence.

I think it is correct to say that RAE* does not perfectly deliver everything we'd like an account of evidence to deliver. In particular, RAE* says that the evidence you have need not supervene on your mental states. This supervenience thesis is one which is undeniably attractive. I've suggested, however, that this supervenience thesis is only

[1] If our starting model is RAE-t rather than RAE, we would add a fourth condition: (iv) every member of $\mathbf{E_t}$ is true.

attractive because it is thought that if evidence supervenes on the mental, then evidence will be the kind of thing that an agent can use, can recognize as evidence, and thus can provide epistemic guidance. But we've seen that this is false (Chapter 4, Section 2.1). Even internal accounts of evidence will fail to provide these features. There may be other ways in which RAE* does not perfectly deliver everything we'd like from an account of evidence, but I believe that it is a promising start.

I have attempted to answer the Evidence Question from a particular theoretical perspective, that of Bayesian Epistemology. Some may see the prominence of this theoretical perspective as a disadvantage. But, as noted in the Introduction, I view a theoretical perspective as necessary to the investigation of the Evidence Question. On its own, the Evidence Question seems too unwieldy to permit of interesting answers and of application to specific cases. By situating the question within the confines of Bayesian Epistemology, much welcomed structure is added to the inquiry, in a principled way.

So, the inquiry has been framed in terms of Bayesian Epistemology. However, the project of Bayesian Epistemology is certainly not complete. Bayesian Epistemology does not come prepackaged with its proper interpretation and correct formulation of its guiding principles. So there is plenty of work to be done in understanding BE itself and how it fits in with the rest of epistemology. So, although the primary goal has been to provide an answer to the Evidence Question, I have also engaged in some of the work of understanding BE, particularly when it was relevant to answering the Evidence Question.

For example, in Chapter 1 I addressed the question of idealization in epistemic theories, in an attempt to clarify how we should understand Bayesian Epistemology. I argued that Bayesian models are best understood as evaluative, anti-procedural epistemic theories. In Chapter 2, I discussed the issue of internalism and externalism in epistemic theorizing and argued that the traditional Bayesian principles exhibit externalist features. This was important in countering the idea that external accounts of evidence are somehow un-Bayesian. This understanding of Bayesian Epistemology allowed me to develop RAE* in Chapters 3 and 4. The account combines process reliabilism from traditional epistemology with the formal models of Bayesian Epistemology, showing one way in which traditional and formal epistemology can usefully interact. With an answer to the Evidence Question in hand, I moved on to consider specific issues. In Chapter 5, I presented an account of how we can model losing evidence, that appealed to RAE* together with a new understanding of the Bayesian principle of conditionalization. I advocated a non-sequential version of conditionalization (hp-*COND*), that instructs one to update on one's total evidence at each time, rather than simply on one's new evidence. Finally, in Chapters 6 and 7, I discussed Dutch Book Arguments. I argued that if we relax the idealization that agents are perfect introspectors—which is rendered more plausible with an account of evidence like RAE*—then there is a way of distinguishing intuitively compelling Dutch Book Arguments from those that have struck many commentators as less compelling. Thus, though the primary goal has been to provide an answer to the Evidence Question, along the way I have discussed issues of broad interest to Bayesian Epistemology.

There are, of course, still questions left to be investigated. There is much work to be done in interpreting and extending the Bayesian framework. As the framework is further extended, there will be further interesting questions that arise about evidence: what it is, how you get it, and what you should do with it once you have it

APPENDIX A

THE BELIEF-FIXING ROLE

In Section 6.4, I explained how allowing evaluation worlds to differ from c-worlds allows us to distinguish errors in the belief-fixing role from other errors. To see how this works, in a bit more detail, consider the following two (partial) credence functions:

$p(A) = 0.2$ $\quad$ $q(\langle q(A) = 0.2\rangle \wedge \langle q(\neg A) = 0.9\rangle) = x > 0$

$p(\neg A) = 0.9$ $\quad$ $q(A) = 0.2$

$q(A) = 0.2, q(\neg A) = 0.8$

The p-function exhibits incoherence in the belief-fixing role. The q-function does not exhibit incoherence in the belief-fixing role. Now, betting according to both functions leads to guaranteed betting losses if we adopt the orthodox account of how to evaluate the payouts of bets. It is trivial to show this with respect to the p-function. We simply offer the agent the following two bets:

Table 20: Bets I and II

Bet I	
A	8
$\neg A$	-2

Bet II	
A	– 9
$\neg A$	1

For the q-function, we first offer:

Table 21: Bet III

Bet III	
$\langle q(A) = 0.2\rangle \wedge \langle q(\neg A) = 0.9\rangle$	$(1 - x)$
$\neg(\langle q(A) = 0.2\rangle \wedge \langle q(\neg A) = 0.9\rangle)$	$-x$

The bettor loses money if $\neg(\langle q(A) = 0.2\rangle \wedge \langle q(\neg A) = 0.9\rangle)$, but wins if $(\langle q(A) = 0.2\rangle \wedge \langle q(\neg A) = 0.9\rangle)$. Thus, if the bettor wins on Bet III, we can simply offer him Bets I and II, to ensure a loss.

However, if we want to only diagnose incoherence in the first sense, then we must evaluate bets differently. One way of doing this is by abstracting away from who the owner of the belief-functions happens to be.[1] Define a bet set, $\mathbf{B_{cr}}$, for function 'cr', as follows: $\{<P, n>: \text{cr}(P) = n\}$. Then stipulate that a bet set condones bets via its ordered pairs: <P, *n*> condones a bet that pays $\pm(1 - n)$ if *P* and $\pm(-n)$ if $\neg P$. The p-function above has a bet set, $\mathbf{B_p} = \{<A, 0.2>, <\neg A, 0.9>\}$. This condones Bets I and II, which together lead to certain loss. This shows that someone with the p-function as their credence function exhibits objectionable incoherence.

The bet set for the q-function is $\mathbf{B_q} = \{\langle \text{q}(A) = 0.2\rangle \wedge \langle \text{q}(\neg A) = 0.9\rangle, x>,<\neg A, 0.2>,<A, 0.8>\}$. This condones Bet III, and Bet I, but it does not condone Bet II. Accordingly, the q-function's bet set does not condone a set of bets that together lead to certain loss. This shows that someone with the q-function as their credence function does not exhibit the objectionable kind of incoherence.

What exactly is going on when we evaluate bets in terms of bet sets? In the example above, the agent's own doxastic state is the object of Bet III, and yet also condones Bets I and II. Bet sets allow us to separate out these two roles, because they

[1] Briggs proposes what she calls the "suppositional test":

> If it is incoherent to believe both *A* and *B*, then it is equally incoherent to suppose both *A* and *B* at the same time and in the same context. But if it is merely Moore-paradoxical to believe both *A* and *B*, then it is perfectly coherent to believe both *A* and *B* at the same time and in the same context. ([2009], p.79-80)

It is difficult to know just how to take this comment. What follows in the text is my attempt to make this clear.

condone bets without revealing the identity of the agent with those credences. Thus, there is no way to tie the condoning of a certain bet with a certain payout from another bet.

APPENDIX B

ASSUMPTIONS FOR REFLECTION DBA

In the body of this chapter I showed that the assumptions in the modified *Reflection* DBA were to blame for the betting losses. However, I showed this after making a simplifying assumption. The assumption was that there is one particular P that the agent learns whenever he learns $\langle \mathrm{cr}_1(A) = r \rangle$. But this is an overly strict way of understanding RA1. To get the modified *Reflection* DBA, all we need is the weaker claim that there is *some* P that is learned whenever the agent learns that $\langle \mathrm{cr}_1(A) = r \rangle$. Though more complex, I will argue that with this assumption, we can still make the same basic argument that the assumption itself encodes inappropriate epistemic behavior.

Consider the case where there is no necessary connection between the truth of these Ps and the truth of $\langle \mathrm{cr}_1(A) = r \rangle$. As before, in such a situation, there is something pathological about such a way of changing beliefs.

Suppose then that there *is* some such necessary connection between the truth of the Ps and the truth of $\langle \mathrm{cr}_1(A) = r \rangle$. If that connection is *a priori*, then the following proposition is *a priori*:

(1) $\Box [(P_1 \vee P_2 \vee \ldots \vee P_n) \leftrightarrow \langle \mathrm{cr}_1(A) = r \rangle]$

If this is *a priori*, then $\mathrm{cr}((P_1 \vee P_2 \vee \ldots \vee P_n) \leftrightarrow \langle \mathrm{cr}_1(A) = r \rangle) = 1$. Further, for each of the P_i $\mathrm{cr}(A|\langle \mathrm{cr}_1(A) = r \rangle \wedge P_i) = r$ as required by RA1. Because (1) is true, it must be that $\mathrm{cr}(A|\langle \mathrm{cr}_1(A) = r \rangle \wedge P_i) = \mathrm{cr}(A|P_i)$. Putting these two facts together, we have:

(2) For each i, $\mathrm{cr}(A|P_i) = r$.

Now, suppose that the P_i are mutually exclusive. If so, then (2) implies that

$\mathrm{cr}(A|P_1 \vee P_2 \vee \ldots \vee P_n) = r$. Suppose now that the P_i completely overlap in the sense that there is P_k such that for all P_i, $(P_k \wedge P_i)$ is equivalent to P_k. If so, then (2) implies that $\mathrm{cr}(A|P_1 \vee P_2 \vee \ldots \vee P_n) = r$. Thus, if the P_i are mutually exclusive or completely overlap, then $\mathrm{cr}(A|P_1 \vee P_2 \vee \ldots \vee P_n) = r$ even though $\mathrm{cr}_1(A|\langle \mathrm{cr}_1(A) = r\rangle) \neq r$ and $\mathrm{cr}((P_1 \vee P_2 \vee \ldots \vee P_n) \leftrightarrow \langle \mathrm{cr}_1(A) = r\rangle) = 1$. This is incoherent.

But what if the P_i are not mutually exclusive or completely overlapping? Then it is possible that $\mathrm{cr}(A|P_1 \vee P_2 \vee \ldots \vee P_n) \neq r$ even though for all i, $\mathrm{cr}(A|P_i) = r$. However, if $\mathrm{cr}(A|P_1 \vee P_2 \vee \ldots \vee P_n) \neq r$, then there must be some P_k and P_j $(j \neq k)$ such that:

(i) P_k and P_j are compatible, and yet

(ii) for no P_m is $(P_m \leftrightarrow (P_k \wedge P_j))$ true.[1]

This means that the P_i are such that though you could learn any one of them, you cannot learn arbitrary conjunctions of them. This is not so odd if the conjunctions are conjunctions of mutually exclusive propositions, but in this case that is not what is going on. The conjunctions that you cannot learn *must* be conjunctions of compatible propositions. This *is* an odd situation.[2]

So, the only scenario where assumption RA1 is defensible is where the proposition

[1] Why is this? Imagine that $\mathrm{cr}(A|P_1) = r$ and that $\mathrm{cr}(A|P_2) = r$, but $\mathrm{cr}(A|P_1 \vee P_2) \neq r$. How could this happen? It can only happen if in the area of overlap between P_1 and P_2, $(P_1 \wedge P_2)$, the ratio of A to that region is not r. For, if the ratio *were* r in the $(P_1 \wedge P_2)$ region, then since $\mathrm{cr}(A|P_1) = r$, it must be that the ratio is r in the region $(P_1 \wedge \neg P_2)$, and the same for the region $(\neg P_1 \wedge P_2)$. Since $(P_1 \wedge P_2)$, $(\neg P_1 \wedge P_2)$, and $(P_1 \wedge \neg P_2)$ partition $(P_1 \vee P_2)$ into three mutually exhaustive parts, and since $\mathrm{cr}(A|(P_1 \wedge P_2)) = r$, $\mathrm{cr}(A|(\neg P_1 \wedge P_2)) = r$, and $\mathrm{cr}(A|(P_1 \wedge \neg P_2)) = r$ it follows that $\mathrm{cr}(A|P_1 \vee P_2) = r$. Summarizing: if the ratio were r in the $(P_1 \wedge P_2)$ region, then $\mathrm{cr}(A|P_1 \vee P_2) = r$. So if it doesn't, then the ratio must not be r in the $(P_1 \wedge P_2)$ region. But if the ratio is not r in that region then $(P_1 \wedge P_2)$ is not something that could be learned, since for every proposition that can be learned, P, $\mathrm{cr}(A|\mathrm{cr}_1(A) = r \wedge P) = r$. So, there must be no P_m such that $(P_m \leftrightarrow (P_k \wedge P_j))$ is true.

(1) $\Box [(P_1 \vee P_2 \vee \ldots \vee P_n) \leftrightarrow \langle \mathrm{cr}_1(A) = r \rangle]$

is *a posteriori*. What would such a proposition be like? Perhaps it is plausible that there would be such a proposition if the P_i described neural configurations that make $\langle \mathrm{cr}_1(A) = r \rangle$ true. And perhaps the agent is just set up so that whenever $\mathrm{cr}_1(A) = r$, he learns this as well as the neural basis for this fact. The odd thing here, however, is that it is not clear why P_i's like this make it reasonable for one to be set up such that $\mathrm{cr}_0(A|\langle \mathrm{cr}_1(A) = r \rangle) \neq \mathrm{cr}_0(A|\langle \mathrm{cr}_1(A) = r \rangle \wedge P_i)$. If the P_i are like this, they don't seem to carry any information that could possibly be relevant to one's credence in A. Accordingly, on any way of understanding RA1, the assumption encodes inappropriate epistemic behavior.

[2] Note that the assumptions in the *COND* DBA do not generate the same oddity. Though the agent can learn one and only one evidence proposition, E, the alternative evidence propositions are incompatible with E. That is not what is going on here.

APPENDIX C

DUTCH BOOKS AND HP-COND

One question that arises is whether or not there is a good DBA against hp-*COND*. First, note that someone who updates according to hp-*COND can* be susceptible to a Dutch Book. To see this consider the scenario below.

At *t1* the agent will either learn *E* or he will not. Further, we suppose that at *t0*, the agent has *L* as evidence, and at *t1* will either lose *L* or will not. Let the agent's credences be distributed as follows:

$\text{hp}(E|L) = d\ (0 < d < 1)$ $\quad [\text{cr}_0(E)]$

$\text{hp}(A|E \wedge L) = n$ $\quad [\text{cr}_0(A|E)]$

$\text{hp}(A|E) = r\ (r \neq n)$ $\quad [\text{cr}_1(A)]$

Now we offer the agent the bets:

Table 22: Bets hp1 and hp2

Bet hp1	
$A \wedge E$	$1 - n$
$\neg A \wedge E$	$-n$
$\neg E$	0

Bet hp2	
E	$(d-1)(r-n)$
$\neg E$	$d(r-n)$

Table 23: Bet hp3

Bet hp3	
A	$r - 1$
$\neg A$	r

We retain LA and PA so have that *E* iff *E* is learned. Bet hp3 is then offered at *t1* iff the agent learns *E* and loses *L*. Notice, however, that this will not result in a book. For consider the scenario where the agent learns *E* but doesn't lose *L*. Since the agent learns *E*, *E* is true, and so the agent wins bet hp2. Bet hp3 is not offered, and so the agent wins money depending on the truth-value of *A*.

To make this a book we must tie the learning of *E* with the losing of *L*. But it is plausible that we can do this. In Chapter 5 I showed how learning certain propositions could lead to strong defeating belief that would expunge propositions from the evidence set. Assume that this is one of those instances. In particular, learning *E* will result in a strong belief in a proposition that defeats *L*, thus resulting in *L* being expunged from the evidence set. Thus, we have two alternatives to consider: (1) at *t1 E* is learned and *L* is lost, and (2) at *t1 E* is not learned and *L* is either lost or not.

If the first alternative obtains then all three bets are offered and condoned. The total payoff is $d(r - n)$. If the second alternative obtains, then hp3 is called off and *E* is false. Bets hp1 and hp2 together yield $d(r - n)$ and we have a book. So, we have the structure of a DBA, and since conditions (i) – (iii) are met, we meet the necessary condition for a *good* DBA.

So, we seem to have a situation where an agent is vulnerable to a Dutch Book simply because she followed the dictates of hp-*COND*. What should be said about this?

First, we are concerned whether or not it is irrational to follow the dictates of hp-*COND*. But the argument is set up in a peculiar way. For there is no reason why following hp-*COND* requires $r \neq n$. In the *COND* DBA we set things up so that there *is* a violation of *COND*. This is why $r \neq n$. But in the DBA just given there is no motivation for this. Given this, following hp-*COND* does not *guarantee* betting losses.

Second, we have assumed here that if *E* is learned, then *E* is true. But RAE* allows situations where false propositions are allowed into the evidence set. It is this possibility, in fact, that makes the need for something like hp-*COND* so pressing. For if false propositions can get *in* to the evidence set, then we would like a way for false

propositions to get *out* of the evidence set. This is part of the motivation for hp-*COND*: adding evidence propositions is not irreversible. But if we drop the assumption that if E is learned, then E is true, then we do not get the DBA. The same basic point can be put in a slightly different way: if all evidence is true, then one can see that there is something undesirable about having propositions expunged from one's evidence set. But absent this assumption, losing evidence is not guaranteed to leave one worse off.

Here, it seems, is one places where RAE* and RAE*-t come significantly apart. If one adopts RAE*-t, then one does not have recourse to the response to the hp-DBA that I just gave. Instead, one must rest one's rejection of the DBA on the first point. Further, one may find it difficult in embracing RAE*-t *and* hp-*COND* with the possibility of losing evidence since every bit of evidence lost is guaranteed to leave one worse off.

There is still an interesting question about hp-*COND*, however. The question is *how* we are to motivate it. We cannot give the simple *COND* DBA for hp-*COND* for at least two reasons. First, that DBA depends on the assumption just rejected (if E is learned, then E is true). Second, even if we could make this assumption, upon learning E funny things can happen to the agent's credence function (e.g., evidence can be *lost*) and so we aren't guaranteed to get a book. How, then, do we motivate hp-*COND*?

We can make some headway in this area if we look at the structure of the update required by hp-*COND*. The rule that is going to be followed, if the agent satisfies hp-*COND* is a generalization of the sq-*COND* rule. It is:

$$\mathrm{cr}_0(A|E) = \mathrm{cr}_1(A|L)$$

where the agent learns all and only E and loses all and only L. When no information is lost, then this simplifies to:

$$cr_0(A|E) = cr_1(A)$$

where the agent learns all and only E. When nothing is learned, then this simplifies to:

$$cr_0(A) = cr_1(A|L)$$

where the agent loses all and only L. This latter pattern is like inverse sq-*COND*.

How does seeing this structure help in motivating hp-*COND*? I think it shows that hp-*COND* describes a certain sort of evidence *responsiveness* that is exactly what we want from an update rule. It says that your assessment of A at *t0* conditional on receiving new evidence, E, should be equal to your assessment of A at *t1* after receiving evidence E, adjusted for the information, L, that you lost at *t1*. Williams ([1980]) argues for sq-*COND* by saying that conditionalizing gives you the closest credences to your previous ones while still taking account of the fact that you learned new information E. A similar sort of justification is available to the advocate of hp-*COND*: following hp-*COND* gives you the closest credences to your previous ones while still taking account of the information that was gained *and lost*.

Whether or not this sort of justification for hp-*COND* is successful, the main point is that one who follows hp-*COND* is not thereby open to a good DBA, since the assumptions needed conflict with the kinds of situations hp-*COND* governs.

BIBLIOGRAPHY

Alston, William. [1989]: *Epistemic Justification Essays in the Theory of Knowledge*, Ithaca: Cornell University Press.

Armendt, B. [1992]: 'Dutch Strategies for Diachronic Rules', *PSA 1992: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **1**, pp. 217-229.

Arntzenius, F. [2003]: 'Some Problems for Conditionalization and Reflection', *Journal of Philosophy*, **100**, pp. 356-70.

Bach, K. [1985]: 'A Rationale for Reliabilism', *The Monist*, **68**, pp. 246–263.

Bacchus, F., Kyberg, H., & Thylos, M. [1988]: 'Against Conditionalization', *Synthese*, **85**, pp. 475-506.

Bales, E. [1971]: 'Act-Utilitarianism: Account of Right-Making Characteristicsor Decision-Making Procedure?', *American Philosophical Quarterly*, **8**, pp. 257-265.

Beebe, J. [2004]: 'The Generality Problem, Statistical Relevance and the Tri-Level Hypothesis', *Noûs*, **38**, pp. 177-195.

Bird, A. [2004]: 'Is Evidence Non-Inferential?', *The Philosophical Quarterly*, **54**, pp. 252-265.

BonJour, L. [1985]: *The Structure of Empirical Knowledge*, Harvard University Press: Cambridge, MA.

Bovens, L. & Hartmann, S. [2003]: *Bayesian Epistemology*, Oxford: New York.

Briggs, R. [2009]: 'Distorted Reflection', *Philosophical Review*, **118**, pp. 59-85.

Christensen, D. [*forthcoming*]: 'Higher-Order Evidence', *Philosophy and Phenomenological Research*.

Christensen, D. [2007]: 'Epistemic Self-Respect', *Proceedings of the Aristotelian Society*, **107**, pp. 319-37.

Christensen, David. [2004]: *Putting Logic In Its Place*, New York: Oxford University Press.

Christensen, D. [1996]: 'Dutch Books Depragmatized: Epistemic Consistency for Partial Believers', *Journal of Philosophy*, **93**, pp. 450-79.

Christensen, D. [1992]: 'Confirmational Holism and Bayesian Epistemology.' *Philosophy of Science*, **59**, pp. 540 - 57.

Christensen, D. [1991]: 'Clever Bookies and Coherent Beliefs', *Philosophical Review*, **100**, pp. 229-47.

Chiappe, D & Vervaeke, J. [1997]: 'Fodor, Cherniak, and the Naturalization of Rationality', *Theory and Psychology*, **7**, pp. 799-821.

Chrisman, M. [2008]: 'Ought to Believe', *The Journal of Philosophy*, **105**, pp. 346-70.

Cherniak, C. [1986]: *Minimal Rationality*, Cambridge, MA: MIT Press.

Cherniak, C. [1983]: 'Rationality and the structure of memory', *Synthese*, **57**, pp. 163-86.

Chugh, D. & Bazerman, M. [2007]: 'Bounded Awareness: What You Fail to See Can Hurt You', *Mind and Society*, **6**, pp. 1-18.

Comesaña, J. [2006]: 'A Well-Founded Solution to the Generality Problem', *Philosophical Studies*, **129**, pp. 27-47.

Conee, E. [2001]: 'Heeding Misleading Evidence', *Philosophical Studies*, **103**, pp. 99-120.

Conee, E. & Feldman, R. [2004]: *Evidentialism*, Oxford: New York.

de Finetti, B. [1937/1980]: 'La Prévision: Ses Lois Logiques, Ses Sources Subjectives', *Annales de l'Institut Henri Poincaré*, **7**, pp. 1-68; translated as 'Foresight: Its Logical Laws, Its Subjective Sources', in H. E. Kyburg, Jr. and H. E. Smokler (eds.), *Studies in Subjective Probability*, Huntington, NY: Krieger.

Döring, F. [1999]: 'Why Bayesian Psychology is Incomplete', *Philosophy of Science*, **66**, Supplement, pp. S379-S389.

Feldman, R. [2001]: 'Voluntary Belief and Epistemic Evaluation', in Matthias Steup (*ed*.), *Knowledge, Truth, and Duty*, New York: Oxford University Press, pp. 77-92.

Feldman, R. [1988]: 'Having Evidence', in David Austin (*ed*.), *Philosophical Analysis*, Dordrecht: Kluwer, pp. 83-104.

Feldman, R. & Conee, E. [1985]: 'Evidentialism', *Philosophical Studies*, **48**, pp. 15–35.

Field, H. [1978]: 'A Note On Jeffrey Conditionalization', *Philosophy of Science*, **45**, pp. 361-367.

Firth, R. [1978]: 'Are Epistemic Concepts Reducible to Ethical Concepts?', in A. Goldman and J. Kim (*eds*.), *Values and Morals*, Dordrecht: D. Reidel, pp. 217-19.

Fitelson, B. [2001]: *Studies in Bayesian Confirmation Theory*, Doctoral Dissertation.

Fodor, J. [1983]: *The Modularity of Mind*, Cambridge, MA: MIT Press.

Folely, R. [2009]: 'Belief, Degrees of Belief, and the Lockean Thesis', in Franz Huber & Christoph Schmidt-Petri (*eds*.), *Degrees of Belief*, Springer.

Frankish, K. [2009]: 'Partial Belief and Flat-Out Belief', in Franz Huber & Christoph Schmidt-Petri (*eds*.), *Degrees of Belief*, Springer.

Garber, D. [1980]: 'Field and Jeffrey Conditionalization.' *Philosophy of Science*, **47**, pp. 142-145.

Goldman, A. [1986]: *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.

Goldman, A. [1979]: 'What is Justified Belief?', in George Pappas (*ed*.), *Justification and Knowledge*, Dordrecht: D. Reidel, pp. 1-23.

Goldman, Alvin. [1978]: 'Epistemics: The Regulative Theory of Cognition', *The Journal of Philosophy*, **10**, pp. 509-523.

Gigerenzer, Gerd. [2000]: *Adaptive Thinking: Rationality in the Real World*, New York: Oxford University Press.

Hájek, A. [2008]: 'Arguments for—or against—Probabilism?', *British Journal for the Philosophy of Science*, **59**, pp. 793-819.

Hájek, A. [2005]: 'Scotching Dutch Books?', *Philosophical Perspectives*, **19**, pp. 139-151.

Harman, Gilbert. [1988]: *Change In View*, Cambridge, MA: MIT Press.

Heller, M. [1995]: 'The Simple Solution to the Problem of Generality', *Noûs*, **29**, pp. 501-515.

Hooker, C. A. [1994]: 'Idealization, Naturalism, and Rationality: Some lessons from Minimal Rationality', *Synthese*, **99**, pp. 181-231.

Horwich, P. [1982]: *Probability and Evidence*, New York: Cambridge University Press.

Howard-Snyder, F. [1997]: 'The Rejection of Objective Consequentialism', *Utilitas*, **9**, pp. 241-248.

Howson, C. [2000]: *Hume's Problem*, Oxford: New York.

Howson, C. [1997]: 'Bayesian Rules of Updating', *Erkenntnis*, **45**, pp. 195-208.

Howson, C. & Urbach, P. [1993]: *Scientific Reasoning: The Bayesian Approach (2nd Edition)*, Peru, IL: Open Court.

Howson, C. & Urbach, P. [1989]: *Scientific Reasoning: The Bayesian Approach*, La Salle, IL: Open Court.

Jeffrey, R. [1965]: *The Logic of Decision*. New York: McGraw-Hill.

Jeffrey, R. [1983]: *The Logic of Decision, 2nd Edition*. Chicago: The University of Chicago Press.

Joyce, J. [2004]: 'Williamson on Evidence and Knowledge', *Philosophical Books*, **45**, pp. 296-305.

Kaplan, M. [*forthcoming*]: 'In Defense of Modest Probabilism', *Synthese*.

Kaplan, M. [2002]: 'Decision Theory and Epistemology', in Paul K. Moser (*ed.*) *The Oxford Handbook of Epistemology*. New York: Oxford University Press.

Kaplan, M. [1996]: *Decision Theory as Philosophy*, New York: Cambridge University Press.

Kelly, T. [2005]: 'The Epistemic Significance of Disagreement', in John Hawthorne and Tamare Gendler Szabo (*eds.*) *Oxford Studies in Epistemology, volume 1*. Oxford: Oxford University Press.

Kim, N. [2009]: 'Sleeping Beauty and shifted Jeffrey conditionalization', *Synthese*, **168**, pp. 295–312.

Kitcher, P. [1992]: 'The Naturalists Return', *The Philosophical Review*, **101**, pp. 53-114.

Klein, P. [2007]: 'How to be an Infinitist about Doxastic Justification', *Philosophical Studies*, **134**, pp. 25-29.

Kornblith, H. [2009]: 'A reliabilist solution to the problem of promiscuous bootstrapping', *Analysis*, **69**, pp. 263-267.

Kornblith, H. [1992]: 'The Laws of Thought', *Philosophy and Phenomenological Research*, **52**, pp. 895-911.

Kotzen, M. [*ms*]: 'A Formal Account of Epistemic Defeat', available at <http://matthewkotzen.net/Matthew%20Kotzen_files/defeatersfinal.pdf>

Kvanvig, J. & Menzel, C. [1990]: 'The basic notion of justification', *Philosophical Studies*, **59**, pp. 235-261.

Kyburg, H. [1961]: *Probability and the Logic of Rational Belief*, Middletown, CT: Wesleyan University Press.

Lagnado, D. & Sloman, S. [2004]: 'Inside and Outside Probability Judgments', in Derek Koehler & Nigel Harvey (*eds*.), *Blackwell Handbook of Judgment and Decision Making*, Malden, MA: Blackwell, pp. 157-176.

Lange, M. [2000]: 'Is Jeffrey Conditionalization Defective by Virtue Of Being Non-Commutative? Remarks on the Sameness of Sensory Experiences', *Synthese*, **123**, pp. 393–403.

Laville, F. [2000]: 'Foundations of Procedural Rationality: Cognitive Limits and Decision Processes', *Economics and Philosophy*, **16**, pp. 117-38.

Lehrer, K. [1990]: *Theory of Knowledge*, Boulder, CO: Westview Press.

Levi, I. [2002]: 'Money Pumps and Diachronic Books', *Philosophy of Science*, **69**, pp. S235-S247.

Levi, I. [1970]: 'Probability and Evidence', in Marshall Swain (*ed*.) *Induction, Acceptance, and Rational Belief*, Dordrecht: D. Reidel.

Levi, I. [1967]: 'Probability Kinematics', *British Journal for the Philosophy of Science,* **18**, pp.197-209.

Lewis, D. [1999]: *Papers in Metaphysics and Epistemology*, New York: Cambridg University Press.

Lewis, D. [1980]: 'A Subjectivist's Guide to Objective Chance', in R. Jeffrey (*ed*.), *Studies in Inductive Logic and Probability*, Volume II, University of California Press.

Littlejohn, C. [*ms*]: 'Evidence and Access', available at < http://claytonlittlejohn.blogspot.com/2009/11/scattered-thoughts-fantl-and-mcgrath-on.html>

Lyons, J. [2009]: *Perception and Basic Beliefs*, New York: Oxford University Press.

Maher, P. [1996]: 'Subjective and Objective Confirmation', *Philosophy of Science*, **63**, pp. 149-174.

Maher, P. [1992]: 'Diachronic Rationality', *Philosophy of Science*, **59**, pp. 120-41.

Makinson, D. [1965]: 'The Paradox of the Preface', *Analysis*, **25**, pp. 205–207.

McGee, V. [1999]: 'An Airtight Dutch Book', *Analysis*, **59**, pp. 257-265.

Meacham, C. [*forthcoming*]: 'Unravelling the Tangled Web: Continuity, Internalism, Uniqueness and Self-Locating Belief' *Oxford Studies in Epistemology, Vol. 3*.

Meacham, C. [2008]: 'Sleeping Beauty and the Dynamics of De Se Beliefs', *Philosophical Studies*, **138**, pp. 245-269.

Meacham, C. [2007]: *Chance and the Dynamics of De Se Beliefs*, Doctoral Dissertation.

Meacham, C. [2005]: 'Three Proposals Regarding a Theory of Chance', *Philosophical Perspectives*, **19**, pp. 281-307.

Milne, P. [2003]: 'Bayesianism vs. Scientific Realism', *Analysis*, **63**, pp. 281-288.

Neta, R. [2008]: 'What Evidence Do You Have?', *British Journal for the Philosophy of Science*, **59**, pp. 89-119.

Osherson, D. N. [1996]: 'Probability Judgment', in E.E. Smith and D. N. Osherson (*eds*.), *Thinking: Invitation to Cognitive Science*, Cambridge, MA: MIT Press, pp. 35-76.

Plantinga, A. [1993]: *Warrant: The Current Debate*, New York: Oxford University Press.

Pollock, J. [2000]: 'Epistemic Norms', in Jaegwon Kim & Ernest Sosa (*eds*.) *Epistemology: An Anthology*, pp. 192-225.

Pollock, J. [1995]: *Cognitive Carpentry*, Cambridge, MA: MIT Press.

Pritchard, D. [2008]: 'Sensitivity, Safety, and Anti-Luck Epistemology', in J. Greco (*ed*.), *The Oxford Handbook of Scepticism*, Oxford: Oxford University Press, pp. 437-483.

Pryor, J. [2001]: 'Highlights of Recent Epistemology', *British Journal for the Philosophy of Science*, **52**, pp. 95-124.

Quine, W. V. O & Ullian, J. S. [1978]: *The Web of Belief, 2nd Edition*, New York: McGraw-Hill.

Ramsey, F. P. [1926/1990]: 'Truth and Probability', in F. P. Ramsey, *Philosophical Papers*, D. H. Mellor (*ed*.), Cambridge: Cambridge University Press.

Roush, S. [2009]: 'Second Guessing: A Self-Help Manual', *Episteme*, **6**, pp. 251-268.

Roush, S. [2006]: *Tracking Truth: Knowledge, Evidence, and Science*, Oxford: Oxford University Press.

Ryan, S. [2003]: 'Doxastic Compatibilism and the Ethics of Belief', *Philosophical Studies*, **114**, pp. 47-79.

Samuels, R., Stich, S., & Faucher, L. [2004]: 'Reason and Rationality', in I. Niiniluoto, et al. (*eds*.), *Handbook of Epistemology*. Dordrecht: Kluwer, pp. 131-179.

Schmitt, F. [1984]: 'Reliability, Objectivity and the Background of Justification', *Australasian Journal of Philosophy*, **62**, pp. 1-15.

Schnieder, W. & Shiffrin, R. M. [1977]: 'Controlled and Automatic Human Information Processing I: Detection, Search, and Attention', *Psychological Review*, **84**, pp. 1-66.

Sedlmeier, P. [1999]: *Improving Statistical Reasoning: Theoretical Models and Practical Implications*, Hillsdale, NJ: Erlbaum.

Shafer, G. [1976]: *A Mathematical Theory of Evidence*, Princeton: Princeton University Press.

Shiffrin, R. M., Dumais, S. T., & Schneider, W. [1981]: 'Characteristics of automatism', in J. B. Long & A. Baddeley (*eds*.) *Attention and performance IX*, Hillsdale, NJ: Erlbaum.

Silins, N. [2005]: 'Deception and Evidence', *Philosophical Perspectives*, **19**, pp. 375-404.

Simon, H. [1986]: 'Rationality in Psychology and Economics', *The Journal of Business*, **59**, pp. S209-S224.

Simons, D. J., & Chabris, C. F. [1999]: 'Gorillas in our midst: Sustained inattentional blindness for dynamic events', *Perception*, **28**, pp. 1059-1074.

Simons, D. J. [2000]: 'Current approaches to change blindness', *Visual Cognition*, **7**, pp. 1-15.

Simons, D. J., Chabris, C. F., Schnur, T., Levin, D. T. [2002]: 'Evidence for preserved representations in change blindness', *Consciousness and Cognition*, **11**, pp. 78-97.

Skyrms, B. [1993]: 'A Mistake in Dynamic Coherence Arguments?', *Philosophy of Science*, **60**, pp. 320-28.

Skyrms, B. [1987]: 'Dynamic Coherence and Probability Kinematics', *Philosophy of Science*, **54**, pp. 1-20.

Skyrms, B. [1975]: *Choice and Chance*, *2nd Edition*, Encino, CA: Dickinson.

Sloman, S. [2002]: 'Two Systems of Reasoning', in T. Gilovich, D. Griffin, & D. Kahneman (*eds*.), *Heuristics and Biases*, New York: Cambridge University Press, pp. 379-396.

Sobel, J. H. [1990]: 'Conditional Probabilities, Conditionalization, and Dutch Books', *PSA 1990: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **1**, pp. 503-515.

Sosa, E. [2000]: 'Skepticism and Contextualism', in J. Tomberlin (*ed*.), *Philosophical Issues*, **10**, pp. 1-18.

Stein, E. [1998]: *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*, New York: Oxford University Press.

Steup, M. [2000]: 'Doxastic Voluntarism and Epistemic Deontology', *Acta Analytica*, **15**, pp. 25-56.

Stich, S. [1990]: *The Fragmentation of Reason*, Cambridge, MA: MIT Press.

Sturgeon, S. [*forthcoming*]: 'Confidence and Coarse-Grained Attitudes', *Oxford Studies in Epistemology*.

Teller, P. [1973]: 'Conditionalization and Observation', *Synthese*, **26**, pp. 218-58.

van Fraassen, Bas. [1984]: 'Belief and the Will', *Journal of Philosophy*, **81**, pp. 235-56.

van Fraassen, Bas. [1995]: 'Belief and the Problem of Ulysses and the Sirens', *Philosophical Studies*, **77**, pp. 7-37.

Vineberg, S. [1997]: 'Dutch Books, Dutch Strategies, and What They Show About Rationality,' *Philosophical Studies*, **86**, pp. 185-201.

Vranas, P. [2007]: 'I Ought, Therefore I Can', *Philosophical Studies*, **136**, pp. 167-216.

Vogel, J. [2000]: 'Reliabilism Leveled', *The Journal of Philosophy*, **97**, pp. 602-623.

Wagner, C. [*forthcoming*]: 'Jeffrey Conditioning and External Bayesianity', *Logic Journal of the IGPL*.

Wagner, C. [2002]: 'Probability Kinematics and Commutativity', *Philosophy of Science*, **69**, pp. 266-278.

Weatherson, B. [*ms*]: 'E ≠ K', available at <http://brian.weatherson.org/EK.pdf>

Wedgewood, R. [2002]: 'Internalism Explained', *Philosophy and Phenomenological Research*, **65**, pp. 349-369.

Weisberg, J. [*forthcoming*]: 'Varieties of Bayesianism', in D. Gabbay, S.Hartmann, & J. Woods (*eds*.), *Handbook of the History of Logic*, Volume 10, Elsevier.

Weisberg, J. [2009]: 'Commutativity or Holism: A Dilemma for Conditionalizers', *British Journal for the Philosophy of Science*, **60**, pp. 793–812.

Williams, P. M. [1980]: 'Bayesian Conditionalisation and the Principle of Minimum Information', *British Journal for the Philosophy of Science*, **31**, pp. 131-44.

Williamson, J. [forthcoming]: 'Objective Bayesianism, Bayesian conditionalisation and voluntarism', *Synthese*.

Williamson, T. [*forthcoming*]: 'Why Epistemology Can't be Operationalized', in Q. Smith, (*ed*.), *Epistemology: New Philosophical Essays*, Oxford University Press.

Williamson, T. [2000]: *Knowledge and Its Limits*, Oxford: Oxford University Press.