UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE VIRTUES OF BAYESIAN EPISTEMOLOGY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MARY FRANCES GWIN
Norman, Oklahoma
2011

THE VIRTUES OF BAYESIAN EPISTEMOLOGY

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PHILOSOPHY

BY

____
Dr. James Hawthorne, Chair

____
Dr. Chris Swoyer

____
Dr. Wayne Riggs

____
Dr. Steve Ellis

____
Dr. Peter Barker

For Ted

How I wish you were here.

Acknowledgements

I would like to thank Dr. James Hawthorne for his help in seeing me through the process of writing my dissertation. Without his help and encouragement, I am not sure that I would have been able to complete my dissertation. Jim is an excellent logician, scholar, and dissertation director. He is also a kind person. I am very lucky to have had the opportunity to work with him. I cannot thank him enough for all that he has given me.

I would also like to thank my friends Patty, Liz, Doren, Murry, Becky, Jim, Rebecca, Mike, Barb, Susanne, Blythe, Eric, Ty, Rosie, Bob, Judy, Lorraine, Paulo, Lawrence, Scott, Kyle, Pat, Carole, Joseph, Ken, Karen, Jerry, Ray, and Daniel. Without their encouragement, none of this would have been possible.

Table of Contents

Abstract

The aim of this dissertation is to address the intersection of two normative epistemologies, Bayesian confirmation theory (BCT) and virtue epistemology (VE). While both are successful in many respects, I argue that the constraints on rational degrees of belief provided by Bayesianism are not enough. VE offers additional constraints on degrees of belief, and plays a salutary role for BCT in the form constraints from background knowledge on the more subjective aspects of Bayesianism. Chapter 1 is an introduction to my project. Chapter 2 presents a brief review of the logic and epistemology of science, Bayesian Confirmation Theory. Chapter 3 presents a recent development in cognitive science, *rational analysis*, which employs a Bayesian approach to understanding human reasoning and bases everyday rationality in formal rationality. Chapter 4 presents historical motivations for turning to virtue epistemology. I argue that given historical considerations virtue epistemology offers a truly novel approach by shifting the focus of analysis from properties of beliefs alone to properties of agents. Chapter 5 presents a development of a particular, reliabilist view in virtue epistemology. Chapter 6 concludes my dissertation. In this chapter I argue that Bayesian Confirmation Theory, as an epistemology of science, should be embedded within virtue epistemology and that at least one familiar problem, the problem of the priors, can be ameliorated.

CHAPTER 1

INTRODUCTION

The purpose of this dissertation is to explore the relationship between two diverse and distinct epistemologies—virtue epistemology and Bayesian epistemology. At first glance, it may seem as there is no relationship between the two. Virtue epistemology is concerned with the attributes of agents that turn justified true belief into knowledge. Bayesian epistemology is concerned with how an idealized agent should reason under uncertainty. What could these two diverse projects possibly have in common?

Most importantly both are agent centered in a way that other theories of justification and knowledge are not. Virtue epistemology focuses on the intellectual character of agents in order to account for how true belief becomes knowledge. Bayesian epistemology focuses on the conditions under which theories are confirmed for idealized agents. Both are, to varying extents, idealizations. With its normative character virtue epistemology is an idealization of real agents, though the intuitions to which the virtue epistemologist appeals are often based on actual human performance. The idealization involved in Bayesian epistemology is also due to its normative character, though the idealization is much more straightforward. The main goal of Bayesian epistemology is to show how we arrive at strongly confirmed theories that are true.

My view of what an epistemology ought to do is similar to what Feldman (2001) calls cooperative naturalism, the view that empirical results from psychology are essential to making progress in answering evaluative questions in epistemology.[1] Evaluative questions are questions in general like, "What is the nature/structure of justification/knowledge?" where the answer given is evaluative and normative. The idea behind cooperative naturalism is that epistemology needs to "cooperate" with science by taking into consideration empirical results in order to move forward in answering these kinds of questions and in putting forth normative theories. Epistemology as traditionally construed is "armchair epistemology" and is concerned with analyzing epistemological concepts and formulating epistemic principles without regard for empirical findings.[2] These concepts and principles are subjected to evaluation by entertaining thought experiments. Epistemologists attempt to locate good ways of reasoning by using every day examples and focusing on methods of reasoning that withstand scrutiny (e.g. withstand skeptical attacks) epistemologists are able to contribute to the improvement of our understanding of the concepts of knowledge and justification.[3] The "armchair epistemologist" is not concerned with how a human agent actually reasons, but only with whether or not the proposed concept withstands the possible counterexample. An epistemological naturalist, on the other hand, thinks that cognitive science has something to say about what our epistemic concepts and principles ought to be because this is how actual human

[1] Richard Feldman, "Naturalized Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epistemology-naturalized/>.

[2] Richard Feldman, "Methodological Naturalism in Epistemology," in *The Blackwell Guide to Epistemology*, John Greco (ed.), Blackwell, Malden, MA (1999), p. 170.

[3] Feldman, "Naturalized Epistemology", op. cit.

agents reason.[4] A cooperative naturalist, then, is someone who thinks that the data from empirical science plays a role in evaluating epistemic principles and concepts, and epistemology should draw from cognitive science.

It has long been an intuition of mine that certain areas of philosophy should look to the sciences for evaluating concepts and even the conceptual analysis itself. For example, it seems to me that metaphysics should look to physics for aid in the conceptual analysis of things like time and causality, philosophy of mind should look to biology and cognitive science for aid in conceptual analysis of mind and consciousness, and epistemology should look to cognitive science and computer science for aid in analysis of concepts like knowledge and justification. That being said, I previously said that my view of the role of epistemology is similar to the cooperative naturalist, but it seems to me that this view is one sided. While epistemology can learn from the empirical results in cognitive science, cognitive science can also draw from epistemology, and indeed recent research in cognitive science has done this. Chater and Oaksford (2007) have put forth a program of *rational analysis*, "a methodology for the rational explanation of empirical data,"[5] employing Bayesian epistemology and logic in order to "rationalize"—give an explanatory account of the reasoning behavior of actual agents which conforms to the laws of probability—reasoning behavior.

[4] Feldman, "Methodological Naturalism in Epistemology," op. cit. Feldman calls this view, "methodological naturalism" here.

[5] Mike Oaksford and Nick Chater, *Bayesian Rationality: The Probabilistic Approach to Human Reasoning.* Oxford University Press: New York, NY (2007), p. 31.

One reason given for their appeal to a probability model (and, Bayesian epistemology) in their program is that Bayesian epistemology is normative.

> "…if we want to explain how it is that people…are able to cope so successfully with their highly uncertain world, the norms of probability provide the beginnings of an answer—to the extent that the mind reasons probabilistically, the normative justifications that imply that this is the 'right' way to reason about uncertainty, and go some way to explaining how it is that the cognitive system deals with uncertainty with a reasonable degree of success."[6]

In other words, Bayesian epistemology does two things for the explanation of the uncertain reasoning and decision making of actual agents: 1) in so far as human agents reason probabilistically, the norms of probability serve as a starting point for explaining the reasoning and decision making behavior of actual agents by giving justifications that imply the correct way to reason and make decisions under uncertainty (e.g. Dutch book arguments) and 2) the norms of probability explain how it is that actual human agents are able to successfully reason and make decisions under uncertain conditions and with limited information.

One problem relevant to epistemological naturalism in general is epistemic normativity, and this problem is a relevant problem for the project of rational analysis. The general problem is this: there is no clear link between the naturalistic criteria that may be set forth for epistemic concepts and normative concepts in epistemology. Or, as Kim (2009) puts it, "If we take the discovery and systematization of such criteria to be the central task of normative epistemology, is there any reason to think that…normative epistemology is a

[6] Nick Chater and Mike Oaksford, "The Probabilistic Mind: Prospects for a Bayesian Cognitive Science" in *The Probabilistic Mind: Prospects for a Bayesian Cognitive Science*, Nick Chater and Mike Oaksford (eds.), Oxford University Press: New York, NY (1999), p. 4.

possible field of inquiry?"[7] The specific problem for rational analysis is: if human agents reason probabilistically, to any extent, and rational analysis can explain how this is, then is there any reason to think that the normative concepts of the "armchair epistemologist" still apply? In other words, if rational analysis can explain human reasoning, mistakes and all, then there doesn't seem to be room left for normative concepts such as justifiedness or knowledge.

One possible response that I offer is Kim's own supervenience response. Just like moral (or, valuational) properties supervene on naturalistic conditions, epistemic properties supervene on non-epistemic, naturalistic conditions.[8] So, if rational analysis can explain reasoning behavior employing Bayesian epistemology, then the normative properties of probability can supervene on the reasoning behavior.

Another possible objection to naturalism in general, and specifically my version of naturalism espoused in this dissertation is from Bealer (cited in Kornblith, 2008). The objection goes that naturalism opposes any kind of inferential rule following because inferential rule following requires acknowledging the force of a priori intuition. Kornblith's response is to say that while a priori considerations are irrelevant to the naturalist, what is relevant to the naturalist is reliability. "…reliable inferential practices are epistemically legitimate; those which are unreliable are not….Rules of inference that tend to

[7] Jaegwon Kim, "What is 'Naturalized Epistemology?" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 547.

[8] Ibid., p. 548.

produce true beliefs in the kinds of environments that human being occupy may fail to live up to a priori standards of cogency, but they are none the worse for that. By the same token, rules of inference that do meet a priori standards may be unworkable in practice or hopelessly mired in problems of computational complexity. These kinds of problems are not in any way ameliorated if the rules do meet a priori standards of cogency. *A priori* standards thus drop out of the picture entirely as simply irrelevant to proper epistemic practice. They fail to bear on the conduct of inquiry."[9] While I am sympathetic to the first part of Kornblith's response, the program of rational analysis seems to contradict the second part.

It may be the case that what makes rules of inference epistemically legitimate is the reliability of the inferential practice. For example, if you start with true premises while employing modus ponens, you will reliably end up with a true conclusion. Or, if a prior probability is a statistical hypothesis and the likelihood is also statistical, then if you use Bayes theorem, you will reliably end with an accurate posterior probability. Or, if we take recent research into consideration, if the probabilities do not matter in terms of extracting information from an environment[10], epistemic or inferential practices that reliably get information may be considered invalid under *a priori* considerations. But, reliability may not be the only standard for evaluating the legitimacy of epistemic

[9] Hilary Kornblith, "Investigating Knowledge Itself" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p.655.

[10] Luciano Floridi, "Logical Fallacies as Informational Shortcuts" in *Synthese* Vol. 167, no. 2 (2009), pp. 317-325.

or inferential practices. Part of both epistemic and inferential practice for human agents involves many other things, such as motivation or cognitive character in general. [11] Reliability alone does not take these other considerations into account. It is merely what virtue epistemologists would call a “success component” to epistemic practice.[12]

It is with these considerations in mind that in this dissertation I argue for the following. First, in Chapter 2 I investigate Bayesian Confirmation Theory, the logic and epistemology of science. In this chapter, I review and prove several versions of Bayes’ theorem, discuss the problem with prior probabilities, and suggest that supplementation from resources outside of the Bayesian purview. Second, in Chapter 3 I discuss a recent development in cognitive science, Rational Analysis, which employs a Bayesian approach to understanding human reasoning. In this chapter, I argue that given the success of Bayesian Confirmation Theory and the unique methodology of rational analysis that bases everyday rationality in formal rationality, contemporary analytic epistemology must go beyond conceptual analysis in order to advance the program of Bayesian Confirmation Theory and epistemology in general. Third, in Chapter 4 I discuss three problems in epistemology—the debate over the structure of justification, the debate over the nature of justification, and the problem of defining knowledge—that go unresolved in the history of epistemology. I argue that given these historical

[11] John Greco, “Knowledge as Credit for True Belief,” in *Intellectual Virtue: Perspectives from Ethics and Epistemology*, Michael DePaul and Linda Zagzebski (eds.), Oxford University Press, New York, NY (2003).

[12] Cf. Linda Zagzebksi, *Virtues of the Mind*, Cambridge University Press, New York, NY (1996).

considerations (i.e. the inability of other, more longstanding theories of knowledge which employ the notion of justified true belief to deal with Gettier problems and skepticism) virtue epistemology offers a truly novel approach by shifting the focus of analysis from properties of beliefs alone to properties of agents. Fourth, in Chapter 5 I develop and discuss a reliabilist version of virtue epistemology. Finally, in Chapter 6 I argue for some connections between Bayesian confirmation theory and virtue epistemology.

CHAPTER 2

BAYESIAN CONFIRMATION THEORY

Introduction

Confirmation theory is the study of the logic and epistemology by which scientific theories may be confirmed, disconfirmed, or refuted by evidence. It is a normative enterprise. Thus, the primary issue treated by confirmation theory is *not* how scientists *actually reason*, but what kinds of inferences based on evidence should count as *correct reasoning*. By analogy, the primary issue treated by deductive logic and the associated epistemology is *not* how people (including mathematicians) *actually reason*. One of the primary motivations for the development of rigorous systems of deductive logic was to put the kind of mathematical reasoning used to justify mathematical theorems on a firm footing. If some common bit of mathematical reasoning turns out to be invalid according to the well-founded rules of deductive logic, so much the worse for that kind of reasoning. And so it should go for a well-founded theory of confirmation. If a specific theory of confirmation really can be established as well-founded, and if it then turns out that some bit of common scientific confirmational reasoning is invalid according to that theory of confirmation, then that bit of scientific reasoning is ill-conceived, regardless of how well regarded it may have become among members of a scientific community. Thus, the goal of confirmation theory is to establish normative standards for the kind of scientific reasoning involved in the evidential evaluation of scientific theories. The primary focus of confirmation theory is to investigate how the evaluation of scientific hypotheses and theories *should proceed* if knowledge (or at least true belief) is to be reliably acquired.

A particular theory of confirmation is a proposal for a logic and epistemology that may, arguably, succeed in the establishment of scientific knowledge. The most prominent theory of confirmation among contemporary philosophers of science is *Bayesian Confirmation Theory* (BCT). The logical foundation for BCT consists of a few axioms of the mathematical theory of probability together with a definition of the notion of conditional probability. From these we can deduce Bayes' theorem, which in its various forms, does the main work of explicating how, according to BCT, evidence bears on epistemic evaluation of empirical hypotheses and theories.

In this chapter we will take a look at the most telling versions of Bayes' theorem. I will describe how Bayes' theorem is supposed to codify the evaluation of scientific hypotheses and theories. We will find that the Bayesian approach has certain specific weakness. However, these soft spots are not fatal flaws. Rather, I will suggest, they call for some kind of supplementation from resources outside the usual Bayesian purview. In later chapters I will make specific suggestion about the kinds of supplementation needed in order for the Bayesian account to do the job of supplying a well-founded logic for the evaluation of scientific hypotheses. This will require showing how the Bayesian logic may be fitted into a larger epistemic context suitable to the goals of scientific enquiry.

Section 1—Theory and Evidence

In the broadest sense, hypotheses and theories are collections of propositions expressed as statements. Propositions represent ways the world might be, and statements are bits of language that express propositions. The things that propositions represent may be described in various ways — e.g., as possible states of affairs, or as possible events. To say that some bit of evidence confirms (or disconfirms) a hypothesis or theory means it raises (or lowers) the probability that the hypothesis or theory is true. Confirmation theory, then, is a theory about the relation between evidence and the collection of propositions that constitute a hypothesis or a theory. For evidence to be one of the relata in this relation, it must be expressible in statements as well. Confirmation theory is not necessarily committed to a realist position regarding propositions, but it does at least require that statements are meaningful sentences of a language that represent ways the world might be. A logic uses statements (declarative sentences of a language) as its basic elements. The relevant logic here is an inductive logic.

When we say that a theory is confirmed by evidence, we simply mean that the evidence supports it – the theory is true in a greater proportion of the possible states of affairs where the evidence statement is true than in the class of possible states of affairs overall (including those possible states of affairs where the evidence statement is true taken together with those where it is false). When an evidence statement is true, the evidence itself is some actual state of affairs, some way the world actually is. Evidence supports a theory via statements that describe the relevant evidential state of affairs. The logic represents how evidential statements may come to support bodies of theoretical statements.

The strongest possible case of theory confirmation is when the relation of confirmational support is a logical entailment. However, in a logical entailment the information contained in the conclusion cannot exceed the information contained in the premises. Thus, confirmation via logical entailment is not particularly useful. Almost all interesting hypotheses go well beyond the available evidence. Furthermore, we often want to draw on scientific hypotheses and theories to make predictions about future events. But for a hypothesis to speak about a future event it must necessarily go beyond the evidence for it, which can only come from events that have already occurred. Thus, the only way to confirm most scientific hypotheses is to step beyond deductive logic. The logic of evidential support for scientific hypotheses and theories must of necessity be some sort of inductive logic, which supports conclusions that have content that goes beyond the content of evidence statements. A Bayesian approach to confirmation provides a probabilistic version of an inductive logic.

Section 2—The Probabilistic Logic of Bayesian Confirmation Theory

Let Pr be a probability function defined over a finite collection of sentences (e.g. in a language for predicate logic with identity) that satisfies the following axioms:

1. $0 \leq \Pr(A) \leq 1$

2. $\Pr(A) = 1$, if A is a tautology

3. $\Pr(A \vee B) = \Pr(A) + \Pr(B)$, if A and B are mutually exclusive (i.e. if ~(A & B) is a tautology).[13]

[13] Andrey Nikolaevich Kolmogorov, *Foundations of the Theory of Probability 2nd English Ed.*, Chelsea Publishing Company: New York, NY (1956), pp. 1-14.

In addition, conditional probability is defined[14] as follows:

$$\Pr(A \mid B) = \Pr(A \,\&\, B) / \Pr(B) \text{ when } \Pr(B) > 0.$$

When Pr(B) = 0 we may either leave the conditional probability undefined or we may define it as automatically equal to 1.

Axiom 1 just says that all probabilities are measures between 0 and 1. Axiom 2 says that if A is logically certain, then its probability is one. Axiom 3 says that when it is logically impossible for the sentences A and B to both be true, the probability of the truth of their disjunction equals the sum of the probabilities of the truths of the individual sentences. The definition of conditional probability just says that the probability of the truth of A given the truth of B is equal to the probability of the truth of the conjunction divided by the probability of the truth of just B. Think of the conditional probability as representing the proportion of cases where A and B are true together among all those cases where B is true. From these three axioms and the definition of conditional probability various versions of Bayes' theorem are immediate consequences. Here is a common version (aka The Simple Form of Bayes' Theorem:

Suppose that Pr(E&B) > 0. Then

$$\Pr(H \mid E\&B) = \frac{\Pr(E \mid H\&B) \times \Pr(H \mid B)}{\Pr(E \mid B)}$$

The proof from the axioms is:

[14] Some treatments axiomatize conditional probability directly, and take unconditional probability to be defined as conditional probability where the condition is a mere tautology. Here I will stay with the usual Kolmogorov approach of taking unconditional probability as primitive. However, there may be good reason to follow Hajek (2003), and take conditional probabilities as basic.

1. $\Pr(H \mid E\&B) = \Pr(H\&(E\&B)) / \Pr(E\&B)$ DCP[15]

2. $= \Pr(E\&(H\&B)) / \Pr(E\&B)$ LE[16]

3. $= \Pr(H\&B) \times \Pr(E \mid H\&B) / \Pr(E\&B)$ DCP

4. $= \Pr(H \mid B) \times \Pr(B) \times \Pr(E \mid H\&B) / [\Pr(E \mid B) \times \Pr(B)]$ DCP

5. $= \Pr(H \mid B) \times \Pr(E \mid H\&B) / \Pr(E \mid B)$

Here H, B, and E may be any sentence at all. To see how BCT employs Bayes' Theorem to explicate the evidential support for scientific hypotheses, let H represent a hypothesis or a theory[17], let E represent an evidence statement, let C represent a statement about experimental conditions, and let K represent a statement of the relevant background knowledge. Then substituting C&K for B into the theorem just derived, we get a version of Bayes' theorem that expresses how evidence obtained under conditions C may support a hypothesis H relative to background knowledge and auxiliary hypotheses contained in K:

$$\Pr(H \mid E\&C\&K) = \frac{\Pr(E \mid H\&C\&K) \times \Pr(H \mid C\&K)}{\Pr(E \mid C \& K)}.$$

What this "simple" formulation expresses is this: given evidence E obtained under experimental conditions C, and relative to background information K, the probability of a hypothesis that accounts for that evidence (at least

[15] Definition of Conditional Probability.

[16] LE is the substitutivity of logically equivalent sentences. It can be proved from the axioms as follows. Suppose that A is logically equivalent to B. Then $(A \vee {\sim}B)$ is a tautology and ${\sim}(A \& {\sim}B)$ is a tautology, so by axiom 1 and then axiom 3 we get $1 = \Pr(A \vee {\sim}B) = \Pr(A) + \Pr({\sim}B)$. And again, from axiom 1 and then axiom 3, since $(B \vee {\sim}B)$ and ${\sim}(B \& {\sim}B)$ are tautologies, we have $1 = \Pr(B \vee {\sim}B) = \Pr(B) + \Pr({\sim}B)$, so $\Pr({\sim}B) = 1 - \Pr(B)$. Substituting this into the previous equation gives $1 = \Pr(A) + \Pr({\sim}B) = \Pr(A) + 1 - \Pr(B)$, so $1 = \Pr(A) + 1 - \Pr(B)$, thus $\Pr(A) = \Pr(B)$.

[17] Henceforth I will usually speak in terms of hypotheses; for my purposes theories are just large, complex hypotheses.

probabilistically accounts for it) is a function of its prior probability (prior to taking the evidence into account), Pr(H | C&K), and the likelihood of the evidence according to the hypothesis, Pr(E | H&C&K), and the simple probability that the evidence would have occurred anyway, regardless of whether H is true, Pr(E | C & K).

In most scientific contexts the likelihood of the evidence on a given hypothesis, Pr(E | H&C&K), is fairly easy to evaluate. The whole point of constructing scientific hypotheses is to account for observable features of the world. And the point of doing experiments is to find cases where the hypothesis H (in conjunction with experimental conditions C and background knowledge K) makes explicit claims about how likely it is that various kinds of evidential events E will result. So, in using Bayes' Theorem to evaluate how strongly the evidence supports H—i.e. to evaluate the posterior probability of H, Pr(H | E&C&K), finding the value of the likelihood factor Pr(E | H&C&K) is generally unproblematic. However, the same cannot be said of the other two factors, the simple probability of the evidence, Pr(E | C&K), and the prior probability of the hypothesis, Pr(H | C&K).

On the face of it the evidence, Pr(E | C&K), is particularly hard to assess. How probable would the evidence E be regardless of what hypothesis is true? One way to try to get around this problem is by decomposing this probability into components via the following theorem of probability theory.

Theorem of Simple Evidential Probability Decomposition

If Pr(C&K) > 0, then Pr(E | C&K) =

$$\Pr(E \mid H\&C\&K) \times \Pr(H \mid C\&K) + \Pr(E \mid {\sim}H\&C\&K) \times \Pr({\sim}H \mid C\&H)$$

The proof for this theorem is:

$$\begin{aligned}
\Pr(E \mid C\&K) &= \Pr(E\&C\&K) / \Pr(C\&K) && \text{DCP} \\
&= \Pr(E\&C\&K\&(H \vee {\sim}H)) / \Pr(C\&K) && \text{LE} \\
&= \Pr((E\&C\&K\&H) \vee (E\&C\&K\&{\sim}H)) / \Pr(C\&K) && \text{LE} \\
&= \Pr(E\&C\&K\&H) + \Pr(E\&C\&K\&{\sim}H) / \Pr(C\&K) && \text{Ax 3} \\
&= \frac{\Pr(E \mid C\&K\&H) \times \Pr(C\&K\&H) + \Pr(E \mid C\&K\&{\sim}H) \times \Pr(E\&C\&K\&H)}{\Pr(C\&K)} \\
&= \Pr(E \mid H\&C\&K) \times \Pr(H \mid C\&K) + \Pr(E \mid {\sim}H\&C\&K) \times \Pr({\sim}H \mid C\&H)
\end{aligned}$$

by DCP and LE.

Putting this result together with the Simple Form of Bayes' Theorem yields the following version of Bayes' Theorem.

Bayes' Theorem for Hypotheses and their Negations:

$$\Pr(H \mid E\&C\&K) = \frac{\Pr(E \mid H\&C\&K) \times \Pr(H \mid C\&K)}{\Pr(E \mid H\&C\&K) \times \Pr(H \mid C\&K) + \Pr(E \mid {\sim}H\&C\&K) \times \Pr({\sim}H \mid C\&H)}$$

This form of Bayes' theorem eliminates the difficulty of evaluating the simple probability of the evidence, Pr(E | C&K), but replaces that difficulty with the need to evaluate the likelihood of the evidence on the negation of hypothesis H, Pr(E | ~H&C&K). In some cases this likelihood is also easy to evaluate. For example, if the hypothesis H says that a particular patient has some specific disease D, then ~H says that disease D is not present. In such cases we may have a blood test with specific likelihoods of turning out “positive for D” (asserted by ‘E’), when H is true (when D is present) and when the test is performed and evaluated in the usual prescribed way (as stated by ‘C’), say, Pr(E | H&C&K) = .99. This is the “true positive rate” for the blood test. The test will also have a specific “false positive rate”—i.e. a specific probability of turning our “positive for D” when ~H is true (when disease D is absent), say, Pr(E | ~H&C&K) = .05 (a 5% false positive rate). (In this kind of case C says that the blood test was administered and evaluated in the standard way, and K contains background knowledge about the accuracy and error rates of this kind of test).

However, in many scientific contexts the value of Pr(E | ~H&C&K) is as difficult to determine as the value of the simple probability of the evidence Pr(E | C&K). For example, *how likely is it that the path of a light ray will bend in the presence of a strong gravitational fields,* E, *if the General Theory of Relativity is false,* ~H? How are we to figure that out? Each specific alternative theory of gravitation may make a different prediction, but ‘~H’ seems to make no specific prediction, and so furnishes no specific likelihood value, Pr(E | ~H&C&K), at all.

A way to try to resolve this problem, which is still effectively the problem of evaluating the simple probability of the evidence, is to expand the simple probability, Pr(E | C&K), further. Let {H_1, H_2, ..., H_n} be any finite collection of alternative hypotheses that are mutually exclusive in the sense that for each pair of them, ~(H_i & H_j), is a tautology. Then an extended version of the Decomposition theorem yields the following result:

Theorem of Simple Evidential Probability Decomposition—continued:

Suppose Pr(C&K) > 0 and that for each pair of hypotheses H_i and H_j under consideration K |= ~(H_i&H_j) (i.e. K logically entail that H_i and H_j are not both true). Then

$$\Pr(E \mid C\&K) = \sum_{j=1}^{n} \Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K) + \Pr(E \mid ({\sim}H_1\&...\&{\sim}H_n)\&C\&K) \times \Pr({\sim}H_1\&...\&{\sim}H_n \mid C\&H);$$

$$\Pr(E \mid C\&K) = \sum_{j=1}^{n} \Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K)$$

if $\Pr(H_1 \vee ... \vee H_n \mid C\&H) = 1$ (i.e. if $\{H_1, H_2, ..., H_n\}$ is an exhaustive set of alternative hypotheses);

$$\Pr(E \mid C\&K) = \sum_{j=1}^{\infty} \Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K)$$

if $\lim_{n\to\infty} \Pr(H_1 \vee ... \vee H_n \mid C\&H) = 1$ (i.e. if $\{H_1, H_2, ..., H_n, ...\}$ is an exhaustive set of alternative hypotheses).

From the above three cases we get the following extended forms of Bayes' Theorem.

Bayes' Theorem for sets of Alternative Hypotheses

Suppose Pr(C&K) > 0 and that for each pair of hypotheses H_i and H_j under consideration K |= ~(H_i&H_j) (i.e. K logically entail that H_i and H_j are not both true). Then

$$\Pr(H_i \mid E\&C\&K) = \frac{\Pr(E \mid H_i\&C\&K) \times \Pr(H_i \mid C\&K)}{\sum_{j=1}^{n}\Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K) + \Pr(E \mid ({\sim}H_1\&...\&{\sim}H_n)\&C\&K) \times \Pr({\sim}H_1\&...\&{\sim}H_n \mid C\&H)}$$

$$\Pr(H_i \mid E\&C\&K) = \frac{\Pr(E \mid H_i\&C\&K) \times \Pr(H_i \mid C\&K)}{\sum_{j=1}^{n}\Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K)}$$

if $\Pr(H_1 \vee ... \vee H_n \mid C\&H) = 1$;

$$\Pr(H_i \mid E\&C\&K) = \frac{\Pr(E \mid H_i\&C\&K) \times \Pr(H_i \mid C\&K)}{\sum_{j=1}^{\infty} \Pr(E \mid H_j\&C\&K) \times \Pr(H_j \mid C\&K)}$$

$$\text{if } \lim_{n\to\infty} \Pr(H_1\vee...\vee H_n \mid C\&H) = 1.$$

The first of these three versions of Bayes' theorem is still troubled by requiring the evaluation of a likelihood of form $\Pr(E \mid (\sim H_1\&...\&\sim H_n)\&C\&K)$, the likelihood of the evidence if none of the explicitly stated hypotheses are true. Such a likelihood will often be as difficult to evaluate as the simple probability of the evidence. That problem may be mitigated if the prior probability that none of the explicitly stated hypotheses are true, $\Pr(\sim H_1\&...\&\sim H_n \mid C\&H)$, is very close to 0. Then the whole term $\Pr(E \mid (\sim H_1\&...\&\sim H_n)\&C\&K) \times \Pr(\sim H_1\&...\&\sim H_n \mid C\&H)$ will be close to 0, and so the posterior probability of hypothesis H_i will approximately be given by the second of these three versions of Bayes' Theorem.

The second of the three versions is unproblematic with regard to the values of likelihoods specified by hypotheses. Its only difficulty is in dealing with values of the prior probabilities, a problem to which we will return below. The third form of Bayes' Theorem above applies when version two fails because only some infinite collection of alternative hypotheses captures all of the theoretical possibilities. The difficulty here is not with the likelihoods, but with actually specifying all of these alternative hypotheses—i.e. the problem is specifying all possible hypotheses that may account for a given domain of phenomena (e.g. all possible theories of gravitation). Still, formally the version of Bayes' theorem is correct. It represents a logical idealization that may only be approximated in real cases (e.g. approximated by the second version when the prior probabilities of hypotheses far out in the sequence are extremely low).

In real cases the best we can often do is to compare the evidential support for those alternative hypotheses we can think of. This pairwise comparison of hypotheses is represented by the Ratio Form of Bayes' Theorem.

Ratio Form of Bayes' Theorem

Suppose $\mathrm{Pr}(C\&K) > 0$. Then (taking the ratio of the simple forms of Bayes' theorem applied to two hypotheses, H_i and H_j. we have,

$$\frac{\mathrm{Pr}(H_j \mid E\&C\&K)}{\mathrm{Pr}(H_i \mid E\&C\&K)} = \frac{\mathrm{Pr}(E \mid H_j\&C\&K)}{\mathrm{Pr}(E \mid H_i\&C\&K)} \text{ x } \frac{\mathrm{Pr}(H_j \mid C\&K)}{\mathrm{Pr}(H_i \mid C\&K)}.$$

In practice scientists usually compare hypotheses pairwise on the evidence. If hypothesis H_i makes the evidence E much more likely than alternative H_j, then the ratio of likelihoods $\mathrm{Pr}(E \mid H_j\&C\&K) / \mathrm{Pr}(E \mid H_i\&C\&K)$ becomes extremely small; then provided that H_i isn't too extremely implausible as compared to H_j before the evidence C&E is taken in to account—i.e. provided that $\mathrm{Pr}(H_j \mid C\&K) / \mathrm{Pr}(H_i \mid C\&K)$ isn't extremely small—the posterior probability of H_j becomes extremely small. This follows from the fact that

$$\mathrm{Pr}(H_j \mid E\&C\&K) < \frac{\mathrm{Pr}(H_j \mid E\&C\&K)}{\mathrm{Pr}(H_i \mid E\&C\&K)} = \frac{\mathrm{Pr}(E \mid H_j\&C\&K)}{\mathrm{Pr}(E \mid H_i\&C\&K)} \text{ x } \frac{\mathrm{Pr}(H_j \mid C\&K)}{\mathrm{Pr}(H_i \mid C\&K)}.$$

Thus, evidence C&E will effectively refute alternative H_j provided that the likelihood ratio $\mathrm{Pr}(E \mid H_j\&C\&K) / \mathrm{Pr}(E \mid H_i\&C\&K)$ approaches 0 and the prior plausibility ratio for the two hypotheses $\mathrm{Pr}(H_i \mid C\&K) / \mathrm{Pr}(H_j \mid C\&K)$ makes H_i not too implausible as compared to H_j. We will delve deeper in to prior plausibilities a bit later. (Here, and in what follows below, we may take the evidence statement ‘C&E’ to represent a large body of accumulated evidence.)

If one particular hypothesis H_i makes the evidence much more likely than every alternative that is at all plausible prior to taking the evidence into account—i.e. if $Pr(E \mid H_j\&C\&K) / Pr(E \mid H_i\&C\&K)$ is extremely small for all alternatives to H_i for which $Pr(H_i \mid C\&K) / Pr(H_j \mid C\&K)$ is not extremely small, then we get the following result.

$$Pr({\sim}H_i \mid E\&C\&K) = \sum_{j \neq i} Pr(H_j \mid E\&C\&K), \text{ so}$$

$$\frac{Pr({\sim}H_i \mid E\&C\&K)}{Pr(H_i \mid E\&C\&K)} = \sum_{j \neq i} \frac{Pr(H_j \mid E\&C\&K)}{Pr(H_i \mid E\&C\&K)}$$

$$= \sum_{j \neq i} \frac{Pr(E \mid H_j\&C\&K)}{Pr(E \mid H_i\&C\&K)} \text{ x } \frac{Pr(H_j \mid C\&K)}{Pr(H_i \mid C\&K)}$$

is extremely small—approaching 0. But notice that

$$Pr(H_i \mid E\&C\&K) = Pr(H_i \mid E\&C\&K) / [Pr(H_i \mid E\&C\&K) + Pr({\sim}H_i \mid E\&C\&K)]$$

$$= 1 / [1 + Pr({\sim}H_i \mid E\&C\&K) / Pr(H_i \mid E\&C\&K)].$$

So as the ratio $Pr({\sim}H_i \mid E\&C\&K) / Pr(H_i \mid E\&C\&K)$ approaches 0, the posterior probability $Pr(H_i \mid E\&C\&K)$ must approach 1, and H_i becomes strongly confirmed by the evidence. Thus, when H_i beats each of its competitors H_j in a contest of likelihood ratios, making each likelihood ratio $Pr(E \mid H_j\&C\&K) / Pr(E \mid H_i\&C\&K)$ approach 0, if H_i is itself not too implausible (prior to the evidence) as compared to those competitors—i.e. if none of the prior probability ratios $Pr(H_i \mid C\&K) / Pr(H_j \mid C\&K)$ are extremely small—then the posterior probability of H_i approaches 1, and H_i becomes strongly confirmed by the evidence C&E.

There is a kind of Bayesian convergence theorem that shows the following. If H_i is in fact true and is evidentially distinct from alternative H_j on each of a sequence of experiments C_1, C_2, ..., C_n (i.e. for each of them C_k, there are possible outcomes O_k for which H_i and H_j differ on the values of likelihoods—i.e., $\Pr(O_k \mid H_j\&C_k\&K) \neq \Pr(O_k \mid H_i\&C_k\&K)$ for some possible outcomes O_k), then it is *very likely* (approaching 1 for large numbers of experiments n) that the actual outcomes E_1, E_2, ..., E_n of the series of experiments will be such as to produce likelihood ratios

$$\frac{\Pr(E_1\&E_2\&...\&E_n \mid H_j\&C_1\&C_2\&...\&C_n\&K)}{\Pr(E_1\&E_2\&...\&E_n \mid H_i\&C_1\&C_2\&...\&C_n\&K)}$$ that approach 0 as n increases.[18]

This Likelihood Ratio Convergence Theorem shows that if we ever discover the true alternative hypothesis (but without knowing it to be true), and if we continually test alternative hypotheses against one another on additional evidence, then eventually the true hypothesis will win the contest of likelihood ratios over all the empirically distinct competitors we ever come to consider. As a practical matter this means that continually testing hypotheses against one another will drive the likelihood ratios for each evidentially distinct alternative, H_j, of the true hypothesis, H_i, towards 0, since:

$$\Pr(H_j \mid (E_1\&C_1)\&...\&(E_1\&C_1)\&K) \quad < \quad \frac{\Pr(H_j \mid (E_1\&C_1)\&...\&(E_1\&C_1)\&K)}{\Pr(H_i \mid (E_1\&C_1)\&...\&(E_1\&C_1)\&K)} \quad =$$

[18] For details and a proof of this theorem see Hawthorne, James, "Inductive Logic", *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/win2011/entries/logic-inductive/>.

$$\frac{Pr(E_1\&E_2\&\ldots\&E_n \mid H_j\&C_1\&C_2\&\ldots\&C_n\&K)}{Pr(E_1\&E_2\&\ldots\&E_n \mid H_j\&C_1\&C_2\&\ldots\&C_n\&K)} \times \frac{Pr(H_j \mid K)}{Pr(H_i \mid K)}$$

provided that the conditions under which the experiments are set up are not biased in favor of either hypothesis over the other—i.e. provided that

$$\frac{Pr(H_j \mid C_1\&C_2\&\ldots\&C_n\&K)}{Pr(H_i \mid C_1\&C_2\&\ldots\&C_n\&K)} = \frac{Pr(H_j \mid K)}{Pr(H_i \mid K)} \ .$$

As this happens, the posterior probabilities of each of H_i's competitors goes to 0.

This still leaves an important role for the prior probabilities to play. For, the above evidential refutation of competitors only applies to those competitors that are evidentially distinguishable from the true hypothesis. The posterior probability of H_i can itself approach 1 only if those alternative hypotheses that are not empirically distinct from the true hypothesis can be laid low by incisive comparative plausibility considerations. That is, when H_j fails to be empirically distinct from H_i, the likelihoods of all evidential outcomes is the same for each of them. Then

$$\frac{Pr(E_1\&E_2\&\ldots\&E_n \mid H_j\&C_1\&C_2\&\ldots\&C_n\&K)}{Pr(E_1\&E_2\&\ldots\&E_n \mid H_j\&C_1\&C_2\&\ldots\&C_n\&K)} = 1 \ \text{so}$$

$$\frac{Pr(H_j \mid (E_1\&C_1)\&\ldots\&(E_1\&C_1)\&K)}{Pr(H_i \mid (E_1\&C_1)\&\ldots\&(E_1\&C_1)\&K)} = \frac{Pr(H_j \mid K)}{Pr(H_i \mid K)} \ .$$

The comparative values of the posterior probabilities depend entirely on whatever plausibility considerations may reasonably be brought to bear in establishing their comparative prior plausibilities.

Even when two hypotheses are empirically distinct, there remains an important role for prior plausibility considerations to play. For, notice that if we consider the true hypothesis H_i to be too extremely implausible as compared to some particular alternative H_j, then it may require a huge amount of evidence to overcome that deficit—to make the product of the comparative plausibilities with the ratio of likelihoods come to favor H_i over H_j by making

$$\frac{Pr(E_1\&E_2\&...\&E_n \mid H_j\&C_1\&C_2\&...\&C_n\&K)}{Pr(E_1\&E_2\&...\&E_n \mid H_j\&C_1\&C_2\&...\&C_n\&K)} \text{ x } \frac{Pr(H_j \mid K)}{Pr(H_i \mid K)} =$$

$$\frac{Pr(H_j \mid (E_1\&C_1)\&...\&(E_1\&C_1)\&K)}{Pr(H_i \mid (E_1\&C_1)\&...\&(E_1\&C_1)\&K)}$$

approach 0. So *prior plausibilities*, presumably based on incisive plausibility arguments, will play an important role even when the alternative hypothesis is empirically distinct from the true hypothesis.

The logic does show that in practice one need not become aware of every possible alternative hypothesis for this Bayesian approach to be effective. Rather, these considerations show that if pairs of hypotheses are continually tested against one another and if among them the true hypothesis H_i eventually comes to be considered and tested, then the true hypothesis H_i will become overwhelmingly more strongly confirmed than all of the available alternative empirically distinct alternative hypotheses. As a result, the disjunction of the true hypothesis with its empirically equivalent rivals becomes very strongly supported by the evidence. If in addition we can bring probative and incisive plausibility considerations to bear in a way that tends to make empirically equivalent rivals of the true hypothesis comparatively much less plausible, then the true hypothesis will become much more strongly confirmed than each of its rivals.

The Bayesian logic of evidential support suggests at least several conditions that need to be satisfied in order to have a reasonable chance of coming to strongly confirm a true hypothesis.

1. We must continually try to come up with new alternative hypotheses—for, the true hypothesis may not be among those we are presently considering.
2. We must try to bring well conceived plausibility consideration to bear in comparing the relative plausibilities of hypotheses against one another, and we must be willing to continually consider new plausibility considerations. That is, the so-called prior probabilities should not be fixed once and for all. As new considerations and new arguments are developed, the "background knowledge" K may evolve and change, and even the weight of the same background K may be reconsidered and reevaluated. Thus, the values of the ratios of prior plausibilities $\Pr(H_j \mid K) / \Pr(H_i \mid K)$ should be open to reevaluation, based on the most telling considerations we can muster.
3. We should continually try to find evidential considerations and experimental arrangements C on which pairs of hypotheses differ with regard to objective likelihoods of possible outcomes. We should then conduct as many of these observations and experiments as we practically can, so that we may bring the resulting likelihood ratios to bear on the

> evidential comparison of one hypothesis to another. These observations and experiments should be conducted in a way that precludes bias in the experimental conditions C that may themselves favor one hypothesis over another—i.e., the conditions C for experiments or observations should be such that $\Pr(H_j \mid C\&K) / \Pr(H_i \mid C\&K) = \Pr(H_j \mid K) / \Pr(H_i \mid K)$.

Condition 1 is a kind of "open-mindedness" condition. Indeed, it goes beyond mere open-mindedness in that it tells us to actively challenge the current best hypotheses with new alternatives. Condition 2 suggests that those new alternatives, as well as the old ones, should be rigorously assessed on the most telling plausibility arguments and considerations we can muster. Suppose that in pursuit of conditions 1 and 2 we in fact succeed finding a true hypothesis and find it to be not hugely implausible as compared to other alternatives. Then, if we avidly pursue fulfillment of condition 3, the Likelihood Ratio Convergence Theorem tells us that enough empirically distinguishing experiments and observations will be very likely (approaching a likelihood 1) to produce evidential outcomes that strongly support the true hypothesis over its empirically distinct rivals. Then, provided that our most probative plausibility considerations also succeed at making the true hypothesis more plausible than its empirically equivalent rivals, the true hypothesis will indeed become highly confirmed. However, this approach to obtaining scientific knowledge via confirmations will never place us in a position to be certain that our best current hypothesis is indeed the true one. This approach may succeed in eventually giving us a highly confirmed hypothesis that is also true, but is likely to do so only if we continually attempt to challenge our current best support hypothesis with new alternatives and with new evidence and with new plausibility considerations. Eventually the best confirmed hypothesis will be the true one (and it will remain well confirmed on future testing), but only if we continue the process of challenging and testing our best current alternative hypotheses and theories. I hope this strikes you as a kind of common-sense scientific epistemology. It is just a rigorously articulated version of how the sciences actually proceed in pursuit of theoretical knowledge. The recommended approach is just a rigorously justified version of scientific epistemological common sense.

Section3—The Epistemic Context of Bayesian Confirmation Theory

It takes a community to make a science. If Bayesian Confirmation Theory is to be an effective account of how theories may legitimately be evaluated, then it should play a normative role in influencing the beliefs and actions of the members of a scientific community. Within the community there will be a range of norms guiding things like how experiments are conducted and recorded, how researchers are trained and vetted, and ethical practices regarding how research is conducted. There are also specific epistemic norms, which often go unarticulated, perhaps because we take them for granted. For example, although it's assumed that a well-trained researcher will not falsify experimental data, and the community has ways of guarding against those researchers who might be tempted to publish false data. Among the kinds of epistemic norms for the community are such things as honesty and accuracy in reporting data, and the practice of reporting on experimental procedures in enough detail to make it possible for experiments to be replicated. Thus, among the things that make good science epistemically reliable are epistemic norms, many of which are instilled in one's training and then become taken for granted.

The logic of BCT, like deductive logic, brings its own norms of internal consistency, independent of any specific beliefs of agents. BCT constrains the beliefs that agents may legitimately hold as a result of the logical impact of evidence claims accepted by the community (subject to community accepted norms for legitimate evidence gathering and reporting – both explicit and implicit), and as a result of publically shared plausibility arguments and assessments.

The most prevalent version of BCT among philosophers takes Bayesian probabilities to represent a measure of ideally coherent belief-strengths of agents. These belief-strength probabilities are usually called *subjective probabilities*; they represent an agent's subjective degrees of belief in the truth of various statements. On this account the posterior probability of a hypothesis is supposed to represent how strongly an agent who is certain of the evidence and background knowledge K should believe the hypothesis H. The idea of taking probabilities to represent degrees of belief comes from Bayesian decision theory, where probabilistic belief-strengths are combined with quantitative desirability strengths, called utilities, to generate net *expected utilities*; those actions having the *highest expected utility* are then regarded as the most desirable actions to take.

To see how the subjective degree of belief idea works as an interpretation of probability, consider how it applies to the standard probabilistic axioms stated earlier. First, by convention we may measure belief strength on a scale from 0 to 1. Second, if we know a statement is a tautology, we assign it the highest degree of belief 1. (If a sentence is a tautology but we are unaware of it, then we are less than ideal agents, and so our belief strengths may fail to meet the axioms). Finally, if we are certain that A and B cannot both be true, then our belief strength in their disjunction is supposed to be the sum of our belief strengths in them individually.

How is this probabilistic measure supposed to apply to my individual belief strengths? My degree of belief that my cat, who is now scratching at the door and mewing, wants to go outside and roll around in the dirt is very high, but I am not certain of this, only very confident that it is true. Indeed, although my confidence is very high, I do not know exactly what probability value to assign it. However, as long as I also assign a low enough probability to the negation of the belief (i.e. that it is not the case that my cat wants to go outside and roll around in the dirt), I will remain in conformance with the norms of the probability rules. My belief strengths will be *probabilistically coherent*.

How can the axioms of probability theory be justified as constraints on ideally rational belief strength? One standard answer comes from Dutch-book arguments. Dutch-book arguments assume that degrees of belief are the fair betting quotients for agents.[19] The Dutch-book theorems show that if your degrees of belief do not conformance to the probability axioms, then you should be willing to enter into a package of bets that guarantees a sure loss. Furthermore,

[19] John Earman, *Bayes or Bust? A Critical Examination of Confirmation Theory.* MIT Press: Cambridge, MA (1992), p. 38.

if your degrees of belief remain in conformance with the probability axioms, then you should be unwilling to enter into a package of sure-loss bets.

Although this degree of belief interpretation of the notion of probability is widely accepted among Bayesians, it turns out to be highly problematic as an interpretation of probabilistic confirmation functions. The most salient problem is the so-called *problem of old evidence*. Suppose that the agent already knows that the coin just tossed tuned up *heads*. On the hypothesis H that the coin is fair (that its chances of coming up heads is 1/2) and under conditions C, that it is tossed in the normal way, what is the agent's proper degree of belief that E, the toss resulted in *heads*? Presumably it is 1, since the agent is already certain that E. But the appropriate Bayesian likelihood for testing H should be Pr(E | H&C) = 1/2, not 1. Bayesian confirmational likelihoods are not about what an agent believes. They represent what the hypothesis *says* or implies about evidential outcomes. Belief does enter in, of course. When the agent believes E, and when E is the totality of the agent's evidence regarding H, then (other things equal) the agent should align her belief-strength with the appropriate confirmational posterior probability, Pr(H | E&C&B), to the extent she can determine its value (or the range within which its value lies). The reason she should do so is that in doing so, given enough evidence, her belief-strength is likely to go towards 0 for false hypotheses and towards 1 for true hypotheses. However, Bayesian confirmation values shouldn't themselves be construed as belief strengths. Rather, they are part of a logical apparatus that may guide belief strengths, much as deductive entailment may help guide the beliefs of agents—i.e. in the deductive case, when premises logically entail a conclusion, if you believe the premises to be true, then you should believe the conclusion to be true, because the truth of the conclusion is guaranteed by the truth of the premises. In the confirmational case, aligning belief strengths with confirmation values has the advantage of (very probably, under appropriate conditions) eventually leading the agent to strongly believe that false hypotheses are false and that the true hypothesis is true.

Section 4—Bayesian Epistemology Meets Traditional Epistemology

So far, I have discussed various aspects of Bayesianism logic and epistemology. Now I want to turn to what I will call *traditional epistemology* and articulate possible relationships between Bayesian epistemology and traditional epistemology. First, I need to make clear what I mean by traditional epistemology. Following Hajek and Hartmann (2010), by *traditional epistemology* I mean a broad class of theories that takes belief to be a primitive notion (rather than a notion that is reducible to something else, like credence), takes both knowledge and belief to be central, binary epistemic concepts, and analyzes these concepts in terms of their properties, grounds, and limits.[20] This is a broad generalization meant to capture a wide variety of views in contemporary epistemology. I will focus on one of these traditional views, virtue epistemology, in Chapter 4, where I will motive the main idea and refine it. Then, in Chapter 5 I will show how virtue epistemology may serve to enhance the Bayesian project. Given the description of traditional epistemology in the previous paragraph, one might think that that Bayesian epistemology and traditional epistemology are at

[20] Alan Hajek and Stephen Hartmann, "Bayesian Epistemology" in *A Companion to Epistemology*, Dancy, J., Sosa, E., Steup, M. (eds.), Wiley: Malden, MA (2010) p. 94.

odds. In some sense they are. First, binary belief is not a primitive notion for Bayesians. Most Bayesians take belief to be analyzable in terms of the notion of degrees of belief, or degrees of confidence, or credences. Binary belief may be explicated as degree-of-belief above some threshold considered suitable or useful for some epistemic context. Consequently, the binary conception of knowledge that is taken to be standard in a wide variety of theories under the banner of traditional epistemology is may *not* be so central a concept of Bayesian epistemology either. Given these differences, how can Bayesian epistemology have any important links with traditional epistemology, and how can traditional epistemology make any use of insights from Bayesian epistemology?

Although on the surface it may seem that there is little in common between the two approaches to epistemology, Hajek and Hartmann (2010) convincingly argue that Bayesian epistemology and traditional epistemology complement each other.[21] One suggestion for complimentarity is that while the Bayesian apparatus provides sound normative constraints on rational belief, these constraints are not exhaustive, and there is much room for supplementation. For example, on the most wide-spread version of Bayesianism, the probability axioms listed earlier provide normative constraints on how degrees of belief should work, and Bayes' theorem gives us further normative constraints on how evidence bears on hypotheses. However, these constraints are not the only constraints needed for belief-strengths to be rational.

[21] Ibid., p. 100.

The main goal of BCT is to provide a logic and epistemology for how we arrive at strongly confirmed theories that are true. As a logic, and to some extent as an epistemology, it is an idealization. Additional normative constraints are surely relevant, not only to belief in general, but also to beliefs that arise in scientific contexts. Traditional epistemology mainly endorses these additional normative constraints via analyses of the concepts of justification and knowledge.

For some kinds of objectivist Bayesians, Bayesian probabilities are determined by compatibility with empirical evidence in terms of chance and frequencies.[22] When degrees of belief cannot be determined by empirical chances or frequencies, degrees of belief are to be determined by some form of a principle of indifference.[23] Subjectivist Bayesians, on the other hand, allow for other factors to influence an agent's assessment of Bayesian probabilities.[24] Sometimes there is agreement among various kinds of Bayesians, sometimes there is not. This is especially crucial in the assessment of prior probabilities for hypotheses. But it seems to me that any serious version of Bayesianism that pretends to be a basis for an epistemology has to only permit rational factors to figure into an agent's assessment of prior probabilities, especially in scientific settings.

[22] Jon Williamson, "Objective Bayesianism with Predicate Languages" in *Synthese* Vol. 163, No. 3 (2008) pp. 341, 343.

[23] Earman, op. cit., p. 197. A generalized version of the Principle of Indifference says that when there are n mutually exclusive and exhaustive outcomes, one should assign the probability of 1/n to each outcome. There are other, more recent variations on the Principle, but for simplicity's sake we can focus on this one for now.

[24] Talbott, William, "Bayesian Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epistemology-bayesian/>.

To illustrate an example of how Bayesians of different kinds assign probabilities, consider an example. Suppose that prior to the toss of a coin (not already known to be fair or not) you are asked what the probability of getting *heads* on the toss is. An objectivist Bayesian may advise that the probability that you should assign to each possible outcome (i.e. *heads* or *tails*) be ½. This is because there are two possible outcomes and there will be only one result, and you do not already have any reason to think that one outcome is more likely than the other. A subjectivist may agree with this recommendation, but for a different reason. A subjectivist like de Finetti (1937, 1980) would advise that any assignment of probability between 0 and 1 to the outcome is permissible, provided that the probability of the opposite outcome is one minus that value. For assigning and any combination of probabilities to the possible outcomes that do not add up to one would lead to a sure-loss betting contract, were such a contract offered.[25] In particular, the agent is certainly permitted to assign a degree-of-belief probability of ½ to *heads* if the agent is equally confident in *heads* and in *tails*. So, although both the subjectivist and the objectivist may agree on the probability assignment in this example, the reasons why the assignment may be ½ are different.

Consider another example. Suppose that you are presented with an urn containing some number of equally sized and weighted marbles, and suppose you

[25] Hacking, Ian, *An Introduction to Probability and Inductive Logic*, Cambridge University Press: New York, NY (2001).p. 169. This is also known as a Dutch book argument. de Finetti is one of the first and most influential Bayesian subjectivists who argues that coherence is the rational constraint on prior probabilities. Also, cf. Talbott's supplement to "Bayesian Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/epistemology-bayesian/supplement2.html>.

are given only the information that some are solid yellow and some are cat's eyes. You have no other information about how many marbles are in the urn or on how many of each kind of marble is in the urn. Suppose you are asked the probability of drawing a cat's eye from the urn. An objectivist Bayesian would usually advise you to assign a prior probability of ½ for the same reasons as in the coin toss example. The only information that you have is that there are two kinds of marbles, and given that you have no other information, this is the rational probability assignment according to the Principle of Indifference. A subjectivist may or may not agree with this. A subjectivist may only require probabilistic coherence for degrees of belief, and may allow other factors to influence your degree-of-belief probability assessment. So, subjective and objective Bayesians may well disagree on probability assignments to outcomes.

However, this divide between Bayesians on the rational constraints for assignments of prior probabilities (i.e. probabilities that are not hypothesis based likelihoods) goes even further. Even within the subjective and objective camps there is lack of widespread agreement on the kinds of rational constraints that prior probabilities should satisfy. For example, the simple Principle of Indifference described above has been modernized by Williamson (2005) as a kind of maximum entropy principle.[26] Williamson's principle has become popular among objectivists, but is not accepted by all. Subjectivists face the same problem of no general consensus on rational constraints for prior probabilities.

[26] Williamson (2005), cited in Williamson (2008), ibid. Here Williamson's principle requires that a degree of belief be representable by a maximally non-committal probability function.

I have presented the issue among Bayesians thus far as a debate between subjectivist and objectivist Bayesians. But I there is a more general problem for BCT having to do with prior probabilities for hypotheses.[27] The *problem of the priors* is most basically a problem of establishing normative standards for assigning prior probabilities to hypotheses. Restricting the problem to the subjectivist/objectivist debate shows that there is not an agreed upon normative standard for assigning prior probabilities among various Bayesian schools, though there are normative standards in play within each school. The objectivist would say that the normative standard is set in terms of fitting to chances or frequencies. The subjectivist Bayesian would say that the normative standard is coherence with the probabilistic axioms (at least), or avoiding possible Dutch book scenarios. So, although objectivists and subjectivists may agree on a prior probability assignment in certain cases, there are numerous cases where they may not agree.

I suggest that Bayesians have much to gain from engaging with the normative constraints developed by traditional epistemologies. These normative constraints may or may not be explicitly representable by the formal apparatus, though I suspect that many could find formal expression, or at least have an impact on the formalism.[28] The next section is devoted to suggesting the kinds of

[27] I am not alone, cf. Elliot Sober (2002) and John Earman (1996) for two very nice discussions of the problem.

[28] At this point, I am not concerned with representing normative constraints formally. However, at least one excellent attempt at doing so has been developed. See Bovens and Hartmann (2003) where they define *reliability* as an endogenous variable (see especially pp. 57-60 for a treatment of eye witness reports).

epistemic norms and constraints that a Bayesian epistemology may find particularly helpful.

Section 5—How May Traditional Epistemic Norms Provide Useful Constraints for a Bayesian Epistemology?

So, what *are* the epistemic norms from traditional epistemology that provide useful constraints on belief for Bayesian confirmation theorists? In this section I will discuss three—reliability, creditability, and motivation. I think these are crucial, though not an exhaustive list. I choose these three because they are important components of a virtue account of knowledge that I will develop in Chapter 4. For now, I count them as normative standards because we take these to be standards that have to be met in successful accounts of knowledge.

First, I need to make clear what I mean by the terms 'norm' or 'normative' and 'constraint'. I understand a norm to be a standard of evaluation rather than simply a rule to be followed. As a positive standard of evaluation, an epistemic norm is something to which we ought to aspire in an epistemic context. For example, one straightforward example of an epistemic norm is that a method should be reliable at getting to the truth if it is to warrant justified belief. The central epistemic aim of belief is truth. To the extent that belief is within our control, we aim believing only what is true. This is especially crucial to scientific inquiry.

In this sense, attention to reliability is normative. It is a standard of positive evaluation. We might call a belief forming process reliable if it usually gets us to the truth, or perhaps if it will eventually get us to the truth. A *constraint* on rationality or on the rational formation of belief is a restriction on what is an acceptable, permissible, or correct approach to achieving belief. So, for example, if something is a normative constraint on belief formation, then it is a restriction on methods of belief formation that upholds (or requires) a positive standard for belief acquisition. It may rule some ways of forming beliefs "acceptable", and may rule others preferable, and may even rule some as obligatory.

I take reliability to be a normative constraint on a method for acquiring rational belief. By this I mean, that if a belief forming process is reliable, then we have a standard of evaluation that constrains what methods are permissible, acceptable, or preferable belief forming methods, given the epistemic goal of arriving at truth. To illustrate, consider a specific reliable belief forming process (that in this case happens to be strongly rule-governed) – the process of forming beliefs via valid deductive reasoning. The reasoning process is reliable because deductive logic is truth preserving. Supposing that we start with true beliefs, a deductively arrived at belief is permissible, and perhaps even obligatory, and its negation is forbidden. By contrast, consider an instance of an unreliable method, a belief forming process that does not reliably get to the truth, like wishful thinking. Even when wishful thinking does produce a true belief, it is only an accident that the agent who formed the belief hit upon the truth. A belief arrived at in this way is not epistemically acceptable or permissible to hold, because it results from an unreliable process.

In Chapters 4 and 5 I will pursue the idea of epistemic norms in greater detail. My main point will be to show how norms from traditional epistemology, especially virtue epistemology, bolster Bayesian epistemology. But before taking that on I want to investigate how the Bayesian approach fits with standards of human cognitive functioning with regard to reason and inference. I turn to these issues in the next chapter.

CHAPTER 3

RATIONAL ANALYSIS: BAYESIAN LOGIC AND THE HUMAN AGENT

Introduction

The Bayesian approach to uncertain inference (and so decision theory as well) provides a model of ideal reasoning in the sense that it can be shown to exhibit important benefits to agents who employ it. I discussed some of these benefits in the previous chapter—e.g., avoiding sure loss situations (so-called "Dutch Books") and evidentially converging on true hypotheses. As is well-known, common reasoning behavior does not always produce valid, or even inductively strong, arguments. Humans are *not* natural Bayesian reasoners, in that we do not literally begin with specific prior probabilities and then determine likelihoods and expectedness in order to ultimately calculate posterior probabilities. We are, however, the kinds of creatures who can (and do) learn how to use the Bayesian logic in various ways.[29] Further, because much human reasoning behavior can be (abstractly and partially) modeled in terms of Bayesian logic, human agents may be characterized as *approximately* Bayesian agents. That is to say that although human agents often do not naturally apply Bayes' Rule (and perhaps they cannot do so) in many circumstances, they may unconsciously *implement* it by way of their cognitive systems, and may on occasion even consciously choose to *employ* it. In this chapter I will explore the idea that human agents are often *approximately* Bayesian by looking into an area

[29] I also think it is questionable that human beings are naturally equipped with full-on deductive logic skills. It may be the case that the rules of inference come more easily to some of us, but as the results of the Wason task noted below show, the logic does not always make itself apparent to most of us.

of psychological research that investigates the connection between human cognition and ideal models of cognitive performance, a research program called *rational analysis*. The main thrust of this chapter is to present the some of the ideas underlying *rational analysis* and to use the findings from an updated version of *rational analysis* to argue for a kind of "soft" naturalist approach to epistemology that employs Bayesian and similar formal models as norms. This chapter proceeds as follows: Section 1 presents a brief sketch of John R. Anderson's original *rational analysis* program, Section 2 presents some problems for this position, Section 3 presents Nick Chater and Mike Oaksford's update to Anderson's original program, and finally Section 4 presents my discussion and assessment of this approach, and its implications for the epistemology of human agents.

Section 1—A Brief Sketch of Anderson's Original Program of Rational Analysis

Rational analysis is a relatively recent development in cognitive science. Its proponents attempt to use Bayesian inference to model a variety of cognitive systems and to use experiments to determine how well actual human performance fits the models. For example, Anderson developed a model to predict memory performance where the goal of memory is assumed to be to provide access to needed information acquired in the past.[30] It is an idealization that represents what memory items are made readily available, given a model of the environment in which the memory is to be retrieved, the value of retrieving the memory, and

[30] John R. Anderson, "Is Human Cognition Adaptive?" in *Behavioral and Brain Sciences*, Vol. 14, (1991), p. 474. I will discuss the nature of goals below, but since I am only giving a preliminary example of an application of rational analysis, I will leave the notion vague, as Anderson does.

the cost of retrieving the memory.[31] On this idealization, a rationally designed information-retrieval system stops retrieving when the probability that the memory is relevant to the current context is low enough that the expected gain (measured in some appropriate way relative to the agent's goals) of retrieving the target memory is less than the cost of retrieving the memory.[32]

Generally a *rational analysis* (an RA) is "an explanation of an aspect of human behavior based on the assumption that it is optimized somehow to the structure of the environment."[33] As applied to human reasoning, the model for this approach is very much like the 'rational man' of economics. As in economics, there is no supposition that the agent consciously chooses its behavior. The only supposition is that the behavior of the agent is due to cognitive processes that solve problems in an optimal way (given resource limitations and costs).[34] This is to say that the processes in the cognitive system are such that, given costs and resource limitations, the behavior of the cognitive system will be optimal (at least in some sense). RA does not assume that humans always *use* optimizing techniques, although they might at times. Rather, RA assumes that humans *implement* optimizing techniques, in the sense that our natural reasoning behavior at least roughly follows the patterns of rational models.

[31] Ibid. Anderson begins with what he calls a "rationally designed information retrieval system." He does not assume that human memory is rationally designed, but the idealized model is rationally designed.

[32] Ibid.

[33] Ibid, p. 471.

[34] Ibid.

The main idea is to explain behavior as well-suited to the goals of a cognitive system of an agent in a given environment.[35]

Before I continue with a sketch of Anderson's program below, I want to discuss both the distinction between *using* an optimization technique and *implementing* an optimization model, and also the issue of what is meant by *goals* in the RA framework. When an agent uses an optimizing technique, e.g., when she tries to maximize her subjective expected utility, she will be aware of the components of the technique, e.g., the subjective probability functions that represent beliefs about available actions and about the agent's environment, a preference ranking that represents an agent's desires, attractions, aversions, etc., and the utility functions that numerically represent such a preference ranking. The choice of an action is a matter of discerning which action has the highest expected utility.[36] However, this is not what real agents do in most decision situations. Rather, in many cases people make decisions that more or less follow the recommendations of the model, without having any specific awareness of the model itself. Rather, they roughly track the model by considering things that the model treats as relevant – their situations, what they value, and how to make the most of the former given the latter. This should not be surprising. The Bayesian

[35] This kind of explanation is not necessarily understood evolutionarily, but might be in some cases. The view may be accompanied by the idea that the mind/brain is composed of cognitive modules for solving specific kinds of problems (survival problems posed in an ancestral environment) for which those modules were selected as 'best available' solutions. But regardless of whether some evolutionary origin may explain how some behavior came to be, the immediate goal of a *rational analysis* is to identify and explain specific kinds of behaviors in terms of optimal problem solving solutions by cognitive systems in a given environment.

[36] Grant Hayden and Stephen Ellis, "Law and Economics After Behavioral Economics", *Kansas Law Review*, Vol. 55, p. 634.

decision model wasn't fully developed until the 20th century, and it was developed as an idealization of good human decision making. What would have been quite surprising is if it had captured all real decision making precisely.

Real agents are usually incapable of using the formal Bayesian model precisely: they are unable to assign probability functions to their systems of belief, assign complete and transitive preference rankings that represent all of their desires, and precisely define their utility functions. Rather, real decision makers try to consider their values and beliefs, and try to best satisfy those values given the constraints they think they face. This sort of rough-and-ready thinking about how to proceed can be characterized by, or at least be refined into, a mathematical model, so it seems fair to say that real agents *implement* such a model, rather than consciously apply it. *Implementation* means that they track the main features of the model, not that they consciously try to follow the model. To say that an agent *implements* a Bayesian decision model means that her natural processes of belief formation, preference ordering, etc., can be captured by the Bayesian machinery. The agent need not appeal to the Bayesian norms in any direct way. Rather, the agent's natural processes may follow Bayesian principles, even if the agent has never been exposed to them (much like a person's eyes *implement* the laws of optics without being aware of those laws). Furthermore, a real agent's implementation of the Bayesian model (and the attendant norms) will, in practice, be only approximate. For one thing, real agents are unlikely to have the computational power needed to fully implement Bayesian Decision theory.

Rather, real agents may be natural systems that approximate a Bayesian decision model.

The fact that real agents approximate the Bayesian Decision model should not be at all surprising. For, decision theorists developed that model as an idealization of actual kinds of human reasoning that are generally recognized as good. Thus, it is no surprise that we are often decent reasoners according to Bayesian norms, since those norms reflect an idealization of best human practice. Likewise, it is no surprise that we follow the norms only approximately – they're an idealization.

Can the usual Bayesian models of belief and decision be extended and improved to more accurately capture the decision behavior of real agents? I think so, but some caution is needed when attempting to extend and improve models of decision and reasoning. To the extent that a model becomes more descriptively accurate, it tends to lose at least some of its normative force. One charge against RA, discussed further below, is that it attempts to maintain the normativity of the model while at the same time making it more descriptively accurate. This runs the risk of "rationalizing" cases where reasoning behavior is poor.[37] A truly accurate psychological model of cognitive decision processes should capture all facets of the workings of cognitive systems, including the kinds of cases where they tend to go wrong or break down, much as an accurate account of the physiology of the visual system should model how visual systems can go wrong or break down, and thus provide defective visual representations of the world. So,

[37] C.f. Branden Fitelson and James Hawthorne, "The Wason Task(s) and the Paradox of Confirmation" in *Philosophical Perspectives,* No. 24, (2010), 207-241.

a psychological model that more accurately captures actual behavior will also capture less than fully rational behavior (perhaps even capturing cases of neurotic behavior). Such a descriptively accurate model must lose some of its normative status, at least in some kinds of cases.

Just as we do not want a descriptive psychological model to treat every instance of human reasoning as fully rational, we must also be careful about selecting appropriate normative models against which to judge human performance. By judging human performances against inappropriate normative models it becomes far too easy to label those performances as irrational. In a specific situation the behavior of real people may not descriptively fit one normative model at all well, but may fit an alternative normative model quite closely. In such cases what counts as rational for RA should be behavior that closely fits the latter model. Thus, although a model that counts as rational in the everyday sense discussed below (such as a bias toward confirmation rather than falsification) may fail to model optimal reasoning behavior for the confirmation of scientific hypotheses, it may in fact both be descriptively accurate for the behavior of real agents, and may be quite rational for the kinds of non-scientific day-to-day situations in which people make most of their judgments. That is, the norm of effectively finding the truth is only one possible desiderata of choice. In many everyday situations we care more about advancing other ends (rather than getting to the truth solely for truth’s sake). Thus, in such situations, behavior that fits an alternative normative model more geared towards achieving the desired ends will count as rational according to the model, and that model will be more suited as a normative standard by which to measure behavior to achieve real human ends.

On the issue of goals, Anderson does not make clear whether the goals of an agent are explicit, representing actual desires, or implicit, representing what an agent should desire given the kind of creature it is and other beliefs and desires it may have. One reason for this, I think, is that Anderson is mostly concerned with refining a model that fits actual behavior. Real agents have both kinds of goals—explicit and implicit. However, Anderson also ascribes goals to subsystems, e.g. to the memory system, in a way that seems independent of the explicit goals of the agent of which the system is part. But can a cognitive system of an agent have its own, independent goals? I think that cognitive systems have their own goals only in a metaphorical sense. That is, when developing a model of an agent with cognitive capacities, goals can be attributed to a cognitive sub-system, but the sub-system has no goals apart from the agent – goal talk for such systems should reduce to talk about the proper functioning of the subsystem in contributing to the agent's abilities. Real agents have their own independent goals; cognitive subsystems (as functionally identified systems) do not.

Above I said that the main idea behind RA is to attempt to explain behavior as well suited to the goals of an agent's cognitive system in a given environment. There are at least two motivations for this type of approach within cognitive science. One is to alleviate what Anderson calls the "induction problem" and the other is to aid with what Anderson calls the "identifiability problem." The induction problem is the problem of inferring the cognitive structure of the mind from the structure of behavior. The identifiability problem is the problem of identifying the true mental structure among multiple, competing proposals which produce the same behavioral consequences.[38] Behavior is observed, but identifying and articulating the underlying mental mechanisms that explain the behavior is difficult at best. *Rational analysis* attempts to alleviate the induction problem by assuming that behavior is optimized to the structure of the environment. The idea is this: if we assume (provisionally) that behavior is optimized in this way and we know the optimal behavior given the environment, then this suggests a cognitive mechanism behind the behavior.[39] To be sure, Anderson's induction problem is a general problem for cognitive psychology—inferring the structure of the 'black box' of the mind from the structure of the behavior. However, if this approach is correct, then a mechanism (for the 'black box') can be suggested given the constraint that we try to optimize our behaviors relative to our environment to achieve our goals. This is not so much a solution to the induction problem, but rather a starting point. The idea is that our behavior is fairly well-matched to our goals. We aren't just following a script or blundering about blindly – rather, we tailor our behavior to our goals. That, at least, is how it seems, so is arguably an appropriate starting place for the explanation of behavior.

Rational analysis attempts to alleviate the so-called *identifiability problem* by giving an abstract explanation of behavior in an environment that bypasses the problem of providing a detailed description of cognitive architecture. The identifiability problem is the problem of identifying the *true mental structure*

[38] Anderson, op. cit.

[39] Ibid.

among multiple, competing proposals which produce the same behavioral consequences. Thus, every proposal that accounts for behavior has to deal with the internal structure of human cognition to some extent. The main issue addressed by RA is whether the cognitive mechanisms are rational decision-theoretic-like structures, or something else—e.g., a set of fast and frugal heuristics (*a la'* Gerd Gigerenzer[40]) that succeeds for evolutionarily important goals like survival and reproductive success in the ancestral environment, but often gets less fundamental things wrong. However, one need not have a detailed account of specific cognitive functions in order to address this issue. That is, if RA is right, then we should be able to get a pretty good first approximation of cognitive structure (as psychological regularity) by assuming that in broad outline it captures (i.e. approximately implements) Bayesian decision theoretic principles. Thus, to the extent that people do what normative Bayesian Decision Theory suggests, it makes sense, as a starting point, to suppose that their mental architecture is roughly isomorphic to that model. After all, normative models of reasoning are 'best practice' idealizations of what we take our reasoning to actually be like.

Anderson offers three reasons for thinking that RA will be more successful than alternative approaches here. First, behavior and the structure of the environment are easier to discern than the details of cognitive architecture. Second, *rational analysis* offers a level of explanation for behavioral data that is more appropriate than behavioral explanations in terms of specific brain

[40] Gerd Gigerenzer, *Rationality for Mortals: How People Cope with Uncertainty*, Oxford University Press: New York, NY (2008).

structures, because the implementing mechanism (the specific ‘hardware’) is not as crucial to the explanation of behavior as the ‘program’ being implemented by the cognitive system. Third, whatever the correct explanation for behavioral data at the level of the implementing mechanism, *rational analysis* (or some alternative at the same level of explanation) is still required to provide an adequate characterization and justification of behavior.[41] I will address each reason in turn.

Anderson’s first reason for the appropriateness of the RA approach seems right, as far as it goes – it really *is* easier to discern behavior and environment than it is to discern precise mechanisms in the brain. In a clinical setting, for example, behavioral data can be recorded and the environment can be controlled precisely. Although specific cognitive mechanisms can be precisely described, they cannot be observed directly in the way that behavior and environment are observed. However, ease of discernment of behavior and environment does not imply anything about the shape of those implementing mental mechanisms. If RA is correct, then some systems in the brain must implement the behaviors in accord with whatever norms are appropriate according to RA.

Anderson’s second reason for the appropriateness of the RA approach is that rational explanation is a more appropriate explanation of behavior than lower level mechanisms. This reflects Anderson’s view that appealing to the ‘gory details’ about behavioral mechanisms is unnecessary: “A rational theory … provides an explanation at a level of abstraction above specific mechanistic proposals. All mechanistic proposals which implement the same rational

[41] Anderson, op. cit.

prescription are equivalent…. [A] rational theory provides a precise characterization and justification of the behavior the mechanistic theory should achieve."[42] Anderson arguably (and admittedly) goes further than he ought at times, eschewing all interest in the structure of the mind:

> The structure driving explanation in a rational theory is that of the environment, which is much easier to observe than the structure of the mind. One might take the view (and I have so argued in overenthusiastic moments, Anderson, in press) that we do not need a mechanistic theory, that a rational theory offers a more appropriate explanatory level for behavioral data. This creates an unnecessary dichotomy between alternative levels of explanation however. It is more reasonable to adopt Marr's (1982) view that a rational theory (which he called the 'computational level') helps define the issues in developing a mechanistic theory (which he called the level of 'algorithm and representation'). In particular, a rational theory provides a precise characterization and justification of the behavior the mechanistic theory should achieve.[43]

I agree with Marr and the non-over-enthusiastic version of Anderson here. Rational explanation specifies the psychological-level mechanisms of the mind while allowing that those psychological processes can be implemented (with more or less precision) by a number of possible lower level mechanisms. Still, the psychological-level of explanation tells us something important about the underlying causal mechanisms—they are constrained by the need to implement the psychological pattern.

Because RA and mechanistic theories are compatible, the third reason is the most compelling reason Anderson gives. If there is a correct mechanistic theory of cognition that describes what Marr (1982, as cited by Anderson) calls

[42] Ibid.

[43] Ibid. The article Anderson cites as "in press" in this passage is the following: Anderson, J. R. "The Place of Cognitive Architectures in a Rational Analysis" in K. Van Len (ed.), *Architectures for Intelligence*, Erlbaum: Hillsdale, NJ (1991).

"the level of algorithm and representation,"[44] then it seems right that, to the extent that the program succeeds, RA will provide a characterization of behavior that will closely fit the correct mechanistic explanation, whatever it is. That is, success at the psychological/computational level provides a constraint on mechanistic theories of cognition. The initial hypothesis, at least, should be that the mechanisms of the mind implement the psychological account, so that the processes and categories of that account will be explained (rather than explained away).

Anderson describes what he calls a program of providing a *rational analysis* of behavior via an inferred cognitive system as a six step process. Here is the program as Anderson describes it:

> 1. Precisely specify the goals of the cognitive system.
> 2. Develop a formal model of the environment to which the cognitive system is adapted.
> 3. Make minimal assumptions about the computational limitations of the cognitive system.
> 4. Derive optimal behavior functions given 1-3.
> 5. Check the empirical evidence to see if the predictions of the behavior are confirmed.
> 6. Repeat to refine the theory.[45]

[44] Ibid, p. 471.

[45] Ibid.

I will briefly explicate each step, and address attendant problems. First, Anderson describes step one so as “to specify the goals being optimized by the cognitive system. Any behavior can be seen as optimizing some imaginable goal. Thus, the mere fact that one can predict a behavior under the assumption of optimality is no evidence for rational analysis. One must motivate the goals to be optimized.”[46] Step one involves identifying a cognitive system in terms of its goals. One must hypothesize the existence of a cognitive system (as a component of an agent) whose function is to accomplish specific kinds of tasks in an environment which will further the goals of an agent. The agent’s attempt to accomplish those goals, and the systems that contribute to it, is ultimately supposed to explain various behaviors of the agent. Since many different goals can be suggested as the reason for a given behavior, reasons must be given for taking the system to have the goals that are proposed by the analysis. In other words, the first step in RA is to specify goals for a cognitive system to achieve, and to argue that the attempt by a cognitive system to accomplish these goals is the reason for the specific kinds of behavior to be explained.

Figuring out what goals an agent’s behaviors are trying to accomplish can be difficult. Later I will provide an example where the goal of the cognitive system in question is to help the agent gain information about the world. No doubt it is often beneficial for an agent to accomplish this kind of information-

[46] Ibid, p. 472. Anderson suggests that “there is the strong constraint that these goals must be relevant to adaptation” (p. 472). He recognizes, however, that "[a] rational theory should stand on its own in accounting for data; it need not be derived from evolutionary considerations” (p. 472) While it is almost certainly true that human cognition has been importantly influenced by natural selection, the ways in which that might happen are complex and take us beyond the scope of this work.

obtaining goal. But the hypothesis that a specific kind of behavior results from a cognitive system's attempts to accomplish a specific goal will often be controversial.[47] Ultimately any such hypotheses are subject to evaluation via the usual standards of the scientific enterprise – e.g., to ask the question, *does this hypothesis provide the best (most plausible) explanation of the behavior, given the totality of the available evidence?*

Step two involves developing a formal model of the environment. There may be numerous ways to specify an environment, and the environmental model being used may tend to favor one optimality hypothesis over another.[48] Thus, a body of evidence may be needed to support the hypotheses about the formal model of the environment that captures important features of the environment in which real cognitive systems operate.[49]

Step three involves identifying and specifying constraints on the cognitive system. Such constraints may prevent it from obtaining globally optimal

[47] What I mean here is that it is controversial to posit a specific goal in order to explain a specific behavior. For example, consider the following. Two speakers are engaging in a conversation. Speaker 1 abruptly turns away from Speaker 2 mid-sentence. There are multiple possible goals that could be posited to explain the observed behavior of Speaker 1. Speaker 1 could have turned away in order to show rudeness to Speaker 2, in order to direct attention to an unexpected, loud noise, in order to direct attention to the unusual behavior of a colleague flailing his arms in the background, and so on. Step two allows us to narrow the scope of possible goals, but even then the hypothesis that a specific kind of behavior results from a specific kind of goal is controversial because not all goals are conscious. There are a myriad of possible unconscious goals that may fit an environment.

[48] For one example, consider the possible specifications of environment in footnote 19 (above). The specifications could be very narrowly prescribed (with small scope) eliminating macro-physical descriptions (e.g. the conversation takes place in the Milky Way Galaxy) in favor of micro-physical descriptions (e.g. Speaker 1 is in brain state α) or very broadly prescribed (with wide scope) excluding micro-physical descriptions in favor of macro-physical descriptions (e.g. Speaker 1 is in the second floor office of Speaker 2 on Earth). So, the way in which one optimality model may favor one hypothesis over another could depend on the level of description of the environment.

[49] Ibid.

solutions. The point of a *rational analysis* is to model behaviors as resulting from a cognitive system that responds appropriately to the structure of the environment in which it must work. RA models need not make strong assumptions about the particular nature of the cognitive system itself. Only minimal constraints on the cognitive system and its resources are identified up front.[50] Indeed, Anderson initially called for only two: the cost on the cognitive system in considering alternatives of action, and the limitations of memory. The cost on a cognitive system can be either physical (e.g. a limit in capacity) or computational (e.g. requires more computational resources than the agent has available, or can make available before action must be taken); these need not be mutually exclusive. Furthermore, the model builds in memory limitations because of the empirical evidence that memory really is limited.

The fourth step involves a normative model of ideal performance. Normative models tell us what cognitive systems should be like. Starting from the assumption that people do pretty much what they ought to be doing, RA proposes that normative models do double duty as (abstract, psychological-level) descriptive models. In cases of reasoning and decision the proponents of *rational analysis* often employ the normative aspects of probability theory and Bayesian decision theory in their accounts of how human beings actually behave. The idea is that this step is supposed to answer the question: *what behaviors maximize utility as defined by the goals set forth in step one*.[51]

[50] Ibid.

[51] Ibid.

Step five involves checking to see if the behavior of cognitive systems can be predicted from the optimal behavior predicted by the model. In other words, the question to be answered is, *do the predictions of the model match the empirical evidence?* This step leads to step six which is reiterating the process to develop a more adequate theory, especially in cases where the predictive model of behavior is wide of the mark.

Section 2—Some Problems for *Rational Analysis*

Rational analysis seems like a promising approach to identifying important features of human cognitive systems, especially those involved in reasoning and decision. From a purely philosophical perspective, if the program succeeds, even in a limited way, we may find that human agents are natural Bayesian reasoners, to at least some degree and in some contexts. If the program fails –we discover that human agents are naturally not very good at Bayesian reasoning – then, to the extent that Bayesian reasoning and decision can be shown to be normatively appealing, so much the worse for natural human inclinations. Courses of study in critical reasoning may become that much more important. From a cognitive science standpoint, the issue is to what extent we naturally follow Bayesian norms in our everyday cognitive processing. We at least seem to have beliefs and other doxastic states, as well as desires, hopes, and other evaluative states. We at least seem to make comparative preference judgments. If the RA program succeeds, we will find that our mental states are governed by something like Bayesian norms. If the program fails then perhaps the best bet is that human psychology is shot through with biases and heuristics that may often go wrong, but were selected for evolutionarily important goals (like survival and reproductive success) in the ancestral environment.[52] Or, perhaps there is no good explanation of behavior at the higher, psychological/computational level of abstraction and all behavior is to be explained at the level of specific lower level cognitive mechanisms.

For most of this section I will look more closely at the specific steps, proposed by Anderson, in the development of a *rational analysis*. I will do so with an eye towards identifying potential problems for *rational analysis* as an approach to the study of cognitive components of human reasoning and decision. If this program is to succeed, these problems must be dealt with. Indeed, I think they can be addressed adequately, to the extent that similar kinds of issue may be addressed in other areas of the epistemology of the sciences. In the last part of this section I will address the assumption of rationality behind *rational analysis*.

The first step suggested by Anderson is to specify the goals to be achieved by a cognitive system. Clearly this brings teleology directly into the research program. The usual philosophical/scientific approach to dealing with the teleology of a natural system (a system not created by an intelligent agent) is to provide a functional analysis of that system (and, the system in this case is the agent as a whole, not some isolated part of the agent). This is what Anderson has in mind with step one. While attributing goals to agents in this way is common,

[52] Cf. Gerd Gigerenzer and Peter M. Todd, *Simple Heuristics That Make Us Smart*, Oxford University Press: New York (1999). In this book Gigerenzer and Todd describe a view of heuristics that we follow in order to make decisions under the constraints of limited time and information that are outcome-equivalent to the decisions suggested by Bayesian Decision Theory.

goal attribution to agents is not uncontroversial.[53] Thus, in the specification of goals of a cognitive system, the idea is presumably that human agents need certain capacities to pursue human ends, and the goal of the cognitive systems is to provide for those capacities. Still, to attempt to specify specific goals is often a somewhat speculative enterprise. Recall that the goals of cognitive systems are not to be merely identified with our conscious goals and desires. So, in attempting to satisfy step one of Anderson's program, there may be some degree of speculation with regard to goals of a cognitive system.

In actual cases of *rational analyses* the goals attributed to a cognitive system tend to be pretty uncontroversial, so often the attribution of goals doesn't require an extensive justification. For example, Anderson posits that the goal (i.e. function) of a memory system in a human is to store information in such a way as to enable the possessor of that memory system to retrieve the most useful items in a fast and efficient way (just as the goal or function of the heart is to pump blood throughout the organism).[54] The implicit goal of information storage with the means of fast and efficient retrieval is plausible enough because such a system would be generally useful for achieving any number of ends. Similar to this is the idea that there are cognitive systems for which the goal (or function) is to gather

[53] For example, consider apparent acts of altruism and associated questions surrounding goal attributions. What is the goal of an apparently altruistic act? Is the goal to bring about a good state of affairs? Is bringing about a good state of affairs good because it brings about pleasant sensations in the agent, or is it good for some other reason? Is there another goal? This set of questions in not exhaustive. There are undoubtedly more possible questions that could arise when attributing goals to behavior that appears altruistic, and none of these are easily or straightforwardly answered.

[54] Speed and efficiency are goals in their own right, but are implicit goals and are often subordinated to the explicit goals of an agent. For example, if an agent has the explicit goal of learning a proof, speed and efficiency of retrieval might be subordinated to accuracy.

information about the world, and other systems whose function is to extract additional hypothetical information via inference. The controversy will not usually be about there being cognitive systems that perform these tasks (i.e. that have this goal), but rather about the way in which these goals/functions are accomplished by the relevant systems. For example, connectionism is a popular mechanistic theory that attempts to explain behavior and intellect in terms of artificial neural nets. Information is stored in the connection strengths between the units that make up the network.[55] Though this is a popular view, it stands in contrast to another popular view, computationalism. Computationalism is the view that cognition resembles digital processing, and that strings of information are produced sequentially according to the instructions of symbolic program.[56] Whether cognition and information storage is a result of the connection strengths between the nodes of the network or is a result of symbol processing is a controversial issue. However, controversial issues such as these are not part of step one of Anderson's process for constructing *rational analyses*. Recall that Anderson thinks that 'nuts and bolts' mechanistic theories in cognitive science are unnecessary.

Step 2 is to develop a formal model of the environment(s) to which the system is adapted. In the case of the memory retrieval system, the model will specify that within a specific kind of environment some sorts of memory items are needed very often, others less often, some very quickly, others not so quickly,

[55] James Garson, "Connectionism", *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2010/entries/connectionism.

[56] Ibid.

some very frequently, etc. The model will place some specific (numerical) constraints on these factors (e.g. some items are needed in fractions of a second, others in a few seconds, and others may not be so time-sensitive). Of course, which sorts of items from memory are needed quickly and frequently will usually differ across environments. These features of the model come from studies of real human performance in common situations. So the model is not simply some *a priori* speculation. In cases of information gathering and reasoning, the models are of typical situations in which human agents perform these tasks, but where it is comparatively easy to figure out best outcomes for human agents in the specific environment, and to figure out what various kinds of information gathering and reasoning systems would do in that situation (i.e. in that environment).

One thing that remains unclear is whether, and when, the various sorts of situations (i.e. environments) in which reasoning tasks occur are to be handled by the same cognitive system or by different cognitive systems. If a proposed system for solving a problem (fulfilling a goal) seems to work well at modeling human performance in some kinds of situations, but fails to accurately model performance in other situations, that may suggest that different systems handle these different kinds of situations. Thus, in the iteration stage (step 6) one may have to subdivide environment models, and distinguish types of situations as importantly different environments that trigger different cognitive systems that have somewhat different goals (e.g. with regard to speed vs. reliability).

In step 3 Anderson suggests that minimal (but reasonable) assumptions be made about the computational limitations of the cognitive system. Anderson suggests that these assumptions be limited those about computational ‘costs’ (in time and energy) and limitation in available memory. However, human agents are not supercomputers (we often have a hard time doing simple computations in our heads, such as those involved in balancing a checkbook). And some kinds of problems in reasoning are known to be computationally intractable even for supercomputers (the so-called NP-hard problems).[57] So, without drawing on assumptions that are too limiting, what is already known about human performance from empirical research, and also what is known about computational complexity, may be brought to bear to place reasonable bounds on computational abilities.[58]

Step 4 is the heart of the approach. Here a specific normative model is employed to derive optimal behaviors for the system in bringing about the identified goals identified in step 1 under the environmental conditions modeled in step 2 subject to the computational constraints identified in step 3. The normative model itself may be subject to controversy—does the model employed as a norm have features that make it truly optimal? Here one must look to the literature on such models (logic, decision theory, artificial intelligence) to find the features of suggested models that are supposed to make them superior ways to accomplish cognitive goals. Given a normative model, one must then apply it to the environment model under the identified computational constraints to derive predicted problem solving behaviors. Actually doing this may not be as

[57] For details see Michael R. Garey and David S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman: New York, NY (1979).

[58] This is not necessarily about conscious computations, but usually about what we can naturally compute via unconscious processes.

straightforward as it sounds, and it may bring in various kinds of background assumptions. For example, one background assumption discussed below is the rarity assumption. The rarity assumption is a principle that is based on the idea that finding out that something is true is more informative than finding something that is not true. The rarity assumption says that the probability of something being the case is rather lower than the probability of that something not being the case.[59] So, looking for confirming instances tends to be more helpfully informative (i.e. more useful in general for the purpose of gaining information) than searching for falsifying instances. However, the rarity assumption is just that—an explicit assumption made by proponents of *rational analysis.* The rarity assumption is controversial because it is not clear that this is really what is going on when confirmation bias (the tendency to search only for confirming instances and to not look for refuting instances) is observed. Furthermore, there are always implicit assumptions in an experimental setup (e.g. that the equipment is functioning properly). So care is especially needed here, and controversy may be unavoidable at this point.

At step 5 one does empirical studies to see whether human subjects behave as predicted by the derived problem solving behaviors in step 4. Here the issue arises as to whether the proposed experiments really get at the cognitive tasks one has attempted to model. For example, if the model predicts a specific optimal behavior, but subjects do not exhibit this behavior, then the model is thought to be

[59] Lance Workman and Will Reader, *Evolutionary Psychology: An Introduction*, Cambridge University Press: New York, NY (2004), p.243.

faulty. This is a common problem faced by experimental psychological studies, and is no worse than usual for such research.

If the observed behaviors do not fit the model, in step 6 the whole process is refined and reiterated. Presumably, goals are reexamined and, if necessary, more carefully specified. The environmental model is refined, perhaps subdividing the model into different environments corresponding to a finer, more detailed specification of cognitive goals. Computational limitations are reexamined and refined. Normative models are fitted more sensitively to the refined goals and environments[60], and appropriate goal directed behaviors are again predicted. Then new experiments are designed and conducted. A major worry here is that one might be able to accommodate any behavior with enough tinkering with the models. So care must be taken not to turn this process into a foregone conclusion.

Finally, one may well wonder whether the supposition of rationality on the parts of human agents that underlies *rational analyses* is warranted. There is ample evidence to show that human performance on reasoning tasks (e.g. the Wason tasks) is flawed. If the rationality of human agents is assumed in this model, then the question to answer is *why do we perform so badly on reasoning tasks?* The proponents of *rational analysis* start their answer by distinguishing between two senses of the term “rational” (though there are obviously more senses of the term as well). Oaksford and Chater (2007), for example, note that there is an informal, everyday sense of ‘rationality’ that we take for granted. *Everyday rationality* is that which concerns beliefs and actions in specified

[60] The nature of the data may direct a researcher to focus on either one of these aspects more or less.

circumstances in daily life.[61] For example, *everyday rationality* is involved when we make decisions and act and when we interpret each other's sentences. When we *do* make mistakes, it is against a backdrop of rational behavior that we take for granted. This rational behavior is exhibited in our regular, everyday patterns of thought and action and is difficult to model in artificial form.[62] They maintain that a quite different sense of the term "rational" refers to the kind of rationality involved in formal reasoning (i.e. mathematical and logical).[63]

I want to defend briefly the notion of *everyday rationality* as distinct from formal rationality because I am cautiously sympathetic to this optimistic outlook on human rationality. On the one hand, humans are very good at both basic and non-basic cognitive tasks. For example, one basic cognitive task that we learn and develop as children is communication. Early on we learn how to communicate our desires in order achieve implicit and explicit goals. One (what I take to be) non-basic cognitive task that we learn (which refines our communication skills) is language. We learn both semantics and syntax from our caregivers at amazing rates. As we age these kinds of tasks become more difficult to learn, and our cognitive abilities to learn may diminish. However once we have the cognitive ability, it is possible for us to retain it until death, even in a diminished capacity.

[61] Oaksford, Mike and Nick Chater, *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*, Oxford University Press: New York, NY, (2007), p. 19.

[62] Nick Chater and Mike Oaksford, "Rational Analysis and Human Cognition" in *Reason and Nature: Essays in the Theory of Rationality*, J. L. Bermudez & A. Millar (eds.), Clarendon Press: Oxford, UK (2002), p. 136. Chater and Oaksford point out that formalizing the common sense of everyday rationality for the development of artificial intelligence is notoriously difficult.

[63] Ibid.

On the other hand, there is ample scientific and anecdotal evidence that humans really do fall short of our normative ideals, especially regarding non-basic cognitive tasks (e.g. reasoning skills). This is why I am *cautiously* sympathetic to the notion of *everyday rationality*. Just because a behavior is common (e.g. bias toward confirmation instances), we cannot assume that it is rational. Furthermore, because we do fall short of our normative ideals so often, the ability to achieve implicit goals does not necessarily indicate rationality in a robust sense of the term.

Section 3—Oaksford and Chater's Extension of the Rational Analysis Program

Oaksford and Chater (2007) have recently extended and improved on Anderson's ideas for RA. The main idea underlying RA is that everyday rationality tends to approach norms of ideal rationality as articulated by various kinds of formal models of rationality. Oaksford and Chater specifically focus on particular aspects of human reasoning (e.g. inductive reasoning), whereas Anderson's methodology was aimed at overall cognition. The Oaksford and Chater version of RA seeks to explain how it is that humans are so successful in most kinds of every day reasoning, but unsuccessful in certain kinds of cases that have been studied. Another aspect of this research aims to find links between legitimate informal reasoning and fallacious informal reasoning. I think this second aspect of the project is important, but here I will focus on the work on explaining reasoning behavior in terms of formal models of rationality.

The starting point for this discussion is the relationship between the everyday rationality described above and formal rationality. First, formal rationality is the rationality of logical and mathematical reasoning. It is defined in terms of formal approaches to deductive and probabilistic reasoning that are paradigms for normative rationality. In cases where human reasoning falls short of the normative standards set forth in these formal models, our epistemic evaluations of these cases are negative. One important version of this kind of reasoning employed by RA is the logic of Bayesian Confirmation Theory described in Chapter 2, where it was introduced as an idealized model of reasoning, a model that applies descriptively only to ideal agents. As idealizations both BCT and formal deductive logic are used as standards by which to judge informal and common sense varieties of reasoning and rationality. In particular cognitive scientists (such as Wason) base their evaluations of human performance in terms of their competence at employing principals of deductive logic.[64]

In Chater and Oaksford's version of RA, the usual, straight-forward applications of formal models of rationality to human reasoning do not always suffice to determine whether that reasoning is any good. One reason for this is the evidence from research programs like Wason's, which seem to show that by the standards of the usual formal models, intelligent human beings tend to perform very poorly on simple reasoning tasks.[65] Chater and Oaksford argue that we have

[64] Wason, P. and P. N. Johnson-Laird. *Psychology of Reasoning*, Harvard University Press, Cambridge, MA (1972), p. 2. Here Wason is borrowing there terms from Chomsky.

[65] Cf. Wason, ibid.

good evidence that people are rational—intelligent enough to flourish in a very complex world. So, if the data from some kinds of rather contrived tests seem to show that almost all human agents tend to be very badly mistaken about decisions and judgments in the test situation, perhaps we should look more closely at auxiliary hypotheses about what counts as a correct judgment in such situations. Perhaps the standard models of formal rationality on which we are drawing are not really appropriate to the kind of situation that the test apparently presents to most people. Perhaps in such test situations as the Wason task we are not actually testing for what we think we're testing for, and perhaps the judgments most people make really are quite reasonable when we look at decision making more broadly, in terms of the real goals of the agent. Human agents do seem to be doing something right. So a central task of cognitive psychology should be to figure out how and why natural reasoning succeeds so well.

The RA approach does not consider everyday rationality to be superior to well established principles of formal rationality. Rather the approach is to use the formal models more subtly—to represent features of human reasoning that are ignored by less sophisticated, blunter applications of formal reasoning models. Here are two examples that are well-known examples of how common reasoning falls short of well-established principles of formal rationality. The purpose of mentioning these examples is to give two cases where common reasoning does not agree with principles of formal rationality if applied too bluntly. The RA approach is to take into consideration the fact that the mistakes are common and to give an account of this fact using formal reasoning models.

The first example is decision theory. Decision theory is a formal model of rationality that treats agents as utility maximizers. However, a well-known "paradox" first suggested by Allais apparently shows that real human agents fail to maximize utility.[66] Typically, this failure is understood as a defect in reasoning. This is to say that because agents fail to act so as to maximize utility, the defect is in *agent* reasoning, not a defect in decision theory and its recommendations. Some have suggested, though, that the frequency of failure to maximize utility shows that decision theory as normally conceived needs to be modified.[67]

The second example is from propositional logic. Propositional logic requires that the following inference from (1) to (2) is valid:

[66] In particular I'm speaking of Allais's paradox. The general form of the problem comes from presenting people with the following two choice situations: (A) a choice between the following two payoff schedules: {(A1: 100% chance of getting $1,000,000) vs. (A2: 89% chance of getting $1,000,000; 10% chance of getting $5,000,000; 1% chance of getting nothing)}; (B) a different choice between the following two payoff schedules {(B1: 89% chance of getting nothing; 11% chance of getting $1,000,000) vs. (B2: 90% chance of getting nothing; 10% chance of getting $5,000,000)}. Most people select option A1 over A2, but select option B2 over B1. This is inconsistent with the usual versions of decision theory, which say that when two options (in a given choice situation) have the same outcome a given percentage of the time, that part of the payoff schedule should be irrelevant to the choice between those options. That is, for choice situation A, the 89% of the time where options A1 and A2 agree on providing a $1,000,000 payoff should not influence the choice of A1 over A2. Similarly, for choice situation B, the 89% of the time where options B1 and B2 agree on a payoff of "nothing" should not influence the choice of B1 over B2. For each choice situation, A and B, only the payoffs that may come about the remaining 11% of the time should distinguish the choice option 1 over option 2 (i.e. the choice of option A1 over A2 for situation A, and choice of B1 over B2 for situation B). However, for this remaining 11% of the time we have the following payoff schedules for the two situations: {(A1: 11% chance of getting $1,000,000) vs. (A2: 10% chance of getting $5,000,000; 1% chance of getting nothing)}; {(B1: 11% chance of getting $1,000,000) vs. (B2: 1% chance of getting nothing; 10% chance of getting $5,000,000)}. Notice that these two portions of the payoff schedules are identical for A1 and B1, and are identical for A2 and B2. So, according to standard decision theory, it is irrational to choose option A1 in choice situation A and yet choose option B2 in choice situation B.

[67] E.g., Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk", *Econometrica* No. 47, (1979), pp. 263-291.

(1) If Mary lives in an apartment, Mary is poor, and if Mary lives in a house, Mary is rich.

(2) Either, if Mary lives in an apartment, Mary is rich, or if Mary lives in a house, Mary is poor.

However, most people regard (1) as highly plausible and (2) as not plausible. So, common intuition is to reject the inference from (1) to (2) (because (2) has to be at least as plausible as (1)).[68]

Because it is so common for people to make incorrect assessments regarding these kinds of reasoning problems, one response is to reject the normative status of formal reasoning. However, anyone who has taught an introductory logic class knows that this leads to the unacceptable result of (what I like to call) logic by consensus. Chater and Oaksford use the following metaphor to describe this view, "…being rational is like a musician being in tune: all that matters is that we reason harmoniously with our fellows."[69] This conclusion, that everyday reasoning should trump formal reasoning is unacceptable because it can actually lead to incoherence. The normativity of formal reasoning is based on showing that violations lead to various sorts of incoherence (e.g. assertions of claims that are jointly contradictory, being subject to Dutch bookable betting behavior, etc.). Indeed, with more careful consideration, most people can be brought to see that commonly identified violations of formal logic are indeed bad reasoning.

[68] Chater and Oaksford (2002), op. cit., p. 140. I adapted an example from Cohen (1981) as cited in Chater and Oaksford. Cohen, like Allais, argues that this shows that every day intuitions must have primacy over formal reasoning.

[69] Ibid.

What RA does is look for cases where apparently bad reasoning that is very common in practice may be rational with respect to a more sophisticated application of a (possibly different) model of formal rationality. The traditional approach in cognitive science and philosophy is to treat rationality as uniform and absolute—as ‘all or nothing’. Theories of rationality are compared with human performance, and when human performance does not meet the normative standard required by the theory, human performance is evaluated negatively (e.g. judge as irrational, illogical, unjustified, etc.). RA rejects this approach as heavy-handed. It sees a proper role for cognitive science in explaining the success of agent’s cognitive processes in achieving goals given the constraints of the environment.[70] In doing this, RA draws on formal, normative principles of rationality to give structure to its descriptive explanations. However, which formal principles are to be employed in the explanation of successful reasoning behavior should be determined by which principles are the most useful in explaining the success of those cognitive processes under examination.[71] That is, successful thought and action in everyday contexts are used as working hypotheses.[72] Furthermore, what counts as successful behavior is behavior that best approximates an optimal solution to a specified problem. The empirical problem is that of explaining why people’s cognitive processes are so successful in achieving their goals (broadly construed). In other words, what behavior counts as rational is not determined *a*

[70] Ibid., p. 145.

[71] Ibid.

[72] While successful thought and action in everyday contexts are assumed, it is also understood that there are cases where people just make mistakes. What counts as ‘successful’ will depend on a lot of factors, not the least of which is the goal that is to be attained.

priori, but determined by taking into consideration goals, environment, and computational limitations of the cognitive system.[73] Let us look briefly at each of these three components.

First, Chater and Oaksford assume that everyday thought and action is centered on achieving goals. I take this to be uncontroversial, provided we construe the notion of a goal appropriately. Decision theory is the paradigmatic formal theory concerned with goals. In economics, which is grounded in decision theory, goals are represented in terms of good outcomes (which need not be material and may even be understood as path dependent), and utilities are associated with achieving them.[74] Because goals can vary for a consumer (e.g. save money, or ensure reliability), utilities can vary with the object or event. Moreover as a theory of rational choice, decision theory treats agents as acting so as to maximize utility. To see how this applies to a real human agent we need to carefully examine and evaluate her goals. But there is a multitude of goals involved in everyday thought and action. We can only evaluate the rationality of a real agent's actions if we can accurately determine her goals.[75] Notice that if the notion of a goal is construed broadly enough, then even unconscious, disinterested inquiry can count as an action aimed at a goal. For example, the passive act of perception *aims at* gaining information from an environment.[76]

[73] Ibid., p. 148.

[74] Ibid.

[75] Ibid.

[76] This construal of the notion of a goal is similar to Irwin's understanding of Aristotle's notion of a goal as a state that is caused by a process that acts as an efficient cause and explains the occurrence of the process. T. Irwin, notes to the *Nicomachean Ethics*, p. 325.

Central to a *rational analysis* of an agent's behavior is the role and structure of the environment. The environment acts as a constraint on finding a successful solution. Everyday rationality is successful relative to an environment and what is achievable there. RA employs formal models to determine what actions would count as successful relative to goals and the environment.[77] In other words, what counts as successful action and thought is only meaningful relative to environmental opportunities and restrictions. RA accommodates this by using a formal model of the structure and role of the environment to help delineate the actions and goals available to the agent.

RA makes minimal assumptions about the cognitive computations of the cognitive system. Rather, it hypothesizes that whatever behavior is optimal, or nearly so, for solving the problem at hand will be implemented by the cognitive system. The reason a suboptimal behavior may be implemented is obvious—real agents in everyday thought and action are computationally limited in their abilities.[78] Thus, the approach is fairly standard for Bayesians and cognitive scientists alike. Minimal assumptions are made about the cognitive limitations of the system because the optimal behavior function derived in the formal model is used as an explanatory tool. However, the full explanation will often involve an account of how the formal model would recommend action X as optimal, but that figuring that out would exceed the computational resources (and time) available

[77] Chater and Oaksford (2002), ibid., pp. 148-149.

[78] Ibid, pp. 150-152.

to the agent; however, action Y runs a close second to X and so a system that settles on Y provides a reasonable course of action.[79]

Section 4—Applying RA to a Specific Case: The Wason Experiments

The Wason experiments are widely known. Here are the details of one version of the Wason selection task setup. The subject is shown four cards on a table, each exhibiting just one of the symbols D, K, 3, 7. The subject is told that each card has a number on one side and a letter on the other side. The subject is then asked to say precisely which cards should be turned over to adequately determine whether the following sentence is true or false: *Every card which has a D on one side has a 3 on the other side*.[80] This conditional sentence is known as "the rule," and is classically understood in the experimental setup expressing the material conditional, if p then q where p corresponds to the response that fulfills the antecedent condition, and q corresponds to the response that fulfills the consequent condition. The correct answer to the task, according to the classical interpretation of the problem as a deductive reasoning problem, is to select the cards labeled D and 7, since turning over these cards could prove the rule true or false. The vast majority of subjects fail to choose this combination of cards.[81]

[79] Ibid, p. 151.

[80] P. Wason and D. Shapiro, "Natural and Contrived Experience in a Reasoning Problem", *Quarterly Journal of Experimental Psychology* No. 23 (1971), p. 63. Here is one of Wason's descriptions of the task. "For example, given the sentence: *Every card which has a D on one side has a 3 on the other side* (and knowledge that each card has a letter on one side and number on the other side), together with four cards showing respectively D, K, 3, 7 hardly any individuals make the correct choice of cards to turn over (D and 7) in order to determine the truth of the sentence."

[81] Ibid, pp. 63-64, Wason and Johnson-Laird (1972), ibid., p. 182. The following table is a recreation of the table printed in Wason and Shapiro (1971), p. 64 and Wason and Johnson-Laird (1972), p. 182. The data is combined from four experiments where p corresponds to D, not-p corresponds to K, q corresponds to 3, not-q corresponds to 7.

One RA approach is to interpret this situation not as a deductive reasoning problem, but rather as an inductive reasoning problem. There are two reasons for reinterpreting the problem in this way. First, Chater and Oaksford note that there has been a tacit acceptance on the part of psychologists of the hypothetico-deductive approach to the confirmation of scientific hypotheses.[82] On the hypothetico-deductive approach the Wason task becomes a problem of attempting to falsify hypotheses—where hypotheses are conjectures to be tested by attempted refutation. This view of philosophy of science has historically been subject to criticism.[83] This criticism makes the interpretation of the selection task subject to doubt. If science does not proceed deductively, why should ordinary reasoning in the selection task proceed deductively? Second, if the Wason task is not interpreted as a deductive inference problem employing falsification, then another model of scientific inference may be better able to account for the results in a way that is consistent with the view that the human subjects are acting rationally.[84] To fill out the details of this account of the Wason task as an inductive inference

Frequency of the selection of cards in four experiments (n = 128)

p and q	59
p	42
p, q and not-q	09
p and not-q	05
other	13

[82] Chater and Oaksford, ibid., p. 155.

[83] The hypothetico-deductive method works like this: observation statements are to be deduced from hypotheses and are subject to verification. If the observation statement is found to be false, this falsified the hypothesis; if the observation statement turns out to be true, this supports the hypothesis. One criticism of the H-D approach is that it cannot account for statistical inference. Another is that the very nature of scientific inference is not deductive, cf. Kuhn (1962).

[84] Chater and Oaksford, ibid., p. 156.

problem, Chater and Oaksford employ a Bayesian model of optimal data selection.[85]

Chater and Oaksford treat the task as one of expected information gain. The problem is to determine which cards to turn over to gain the most information regarding the truth of the rule. Information gain is defined as the difference between uncertainty before receiving the data from the other side of a card and the uncertainty after receiving that data. Uncertainty is measured using Shannon-Wiener information which employs both prior and posterior probabilities for information calculations. Bayes' theorem is employed to calculate the posterior probabilities from prior probabilities and likelihoods of the data on the rule and on its negation. The calculations also include a 'rarity assumption' as a default. The rarity assumption is that only a small number of things in the world satisfy the antecedent or the consequent.[86]

[85] Ibid, pp. 156-157.

[86] Ibid, p. 157. Chater and Oaksford define information gain in the following way in Oaksford and Chater *Bayesian Rationality*, pp. 170-171. First, Oaksford and Chater describe the participant's job in the selection task as a task that requires a choice between two hypotheses, M_D and M_I (1994, "A Rational Analysis of the selection task as Optimal Data Selection" in Psychological Review, 101, 608-631). In M_D, the hypothesis is that there is a dependency relation between the antecedent p and the consequent q in the conditional rule, 'if p then q'. In $M_{I,}$ the hypothesis is that the occurrence of *p*s and *q*s are independent. On their characterization, participants want to know which hypothesis accurately describes the relationship between the numbers and letters on the cards, and their task is to choose the data that will give them the most information in order to make this choice. Oaksford and Chater argued that people want to choose the data that will reduce their uncertainty about which of these two hypotheses to select, and the most informative data will be such that it reduces uncertainty the most. Uncertainty is measured in the following way, where $P(M_i)$ is the prior probability that either M_D or M_I accurately describes the relationship between letters and numbers on the cards in the task.

$$I(M_i) = \sum_i P(M_i)\log_2[1/P(M_i)]$$

The uncertainty is at a maximum when $P(M_D) = P(M_I) = .5$, or in other words when M_D and M_I are equally likely. This is the prior uncertainty before turning over a card (data selection).

The net result of this approach is this: by measuring uncertainty in this way, the order of expected information gain matches the empirical results of the Wason task, and this approach explains the data as a rational inductive strategy. The order of expected information (E) and information gain (I_g) for the cards D, 3, 7, and K is $E(I_g(D))> E(I_g(3))> E(I_g(7))> E(I_g(K))$. This corresponds to the observations from the Wason task.[87] In other words, the goal is to gain

To determine information gain from turning over a card requires figuring out uncertainty after data selection D. Uncertainty after data selection is calculated in the following way.

$$I(M_i|D) = \sum_i P(M_i|D)\log_2[1/P(M_i|D)]$$

To determine posterior uncertainty after data selection, Bayes' theorem is employed.

$$P(M_i|D) = P(M_i)P(D|M_i)/\sum_j P(M_j)P(D| M_i)$$

To determine the likelihoods, $P(D|M_D)$ or $P(D|M_I)$ requires a little explanation. Consider a case where someone is thinking about turning over the p card because they think there is a q on the back. The probability of finding a q given M_D is just $P(q|p)$. This is because there is a (at the very least) a perceived relation of dependence between p and q. The probability of finding a q given M_I, however, is $P(q|p \& M_I) = P(q|M_I)$. The probability that the dependence model is true given that someone does find a q on the other side of a p card is $P(M_D|p \& q)$. It follows that $P(M_I| p \& q) = 1 - P(M_D| p \& q)$.

[87] Ibid, p. 158. Expected information gain is calculated in the following way (ibid, pp. 171-174). In general, information (I_g) is the difference between the initial uncertainty regarding which hypothesis is true and the uncertainty *after* data is received

$$I_g = I(M_i) – I(M_i|D).$$

To determine uncertainty associated with finding a q on the back of a p card (p_q)

$$I_g (p_q) = I(M_i) – I(M_i|p_q).$$

$I_g (p_{\sim q})$ is calculated in the same way. In the selection task, participants do not actually turn a card over, so the response (choice of card) is based on the expected information gain from turning a card. To determine this, calculating the probabilities of data outcomes and how much one would *learn* from the outcomes is required. First, the probabilities need to be calculated over both models

$$P(q|p) = P(M_D)P(q|p \& M_D) + P(M_I)P(q|p \& M_I) \text{ and } P(\sim q|p) = 1 - P(q|p).$$

The posterior information gain values are weighted by both $P(q|p)$ and $P(\sim q|p)$ to give the expected uncertainty associated with turning over the p card ($EI(p)$)

$$EI(p) = P(q|p)I_g (p_q) + P(\sim q|p)I_g (p_{\sim q}).$$

information from the environment (i.e. from the D, K, 3, and 7 cards). Because these kinds of objects are rare (i.e. cards with 'D,' 'K,' '3,' and '7' printed on them are rare), the expected amount of information gain from looking for a falsifying instance is calculated as being low, because the probability of finding the falsifier (i.e. 'D' printed on the other side of the 7 card) is considered low.[88] The rational inductive strategy in this case is to look for confirming instances before looking for falsifying instances because that approach is expected to extract more information from the environment.

To say that looking for confirming instances gives more information than falsifying ones may sound counterintuitive. Falsifying instances are, after all, maximally informative. However, because such cards as these are rare and information gain is the goal, the explanation is that confirming instances provide more information *within this sort of environment.* The intuitive example that Chater and Oaksford appeal to is this:

The expected information gain (EI_g) associated with turning over the p card is then

$$EI_g(p) = I(M_i) - EI(p).$$

[88] Ibid. Oaksford and Chater (2007) assume that card choice is competitive, meaning that the information gain associated with turning over each individual card varies, and the choice of whether or not to turn over a particular card x contends with the choices to turn over other individual cards y, z, etc. This assumption allows for Oaksford and Chater to scale information gain associated with each card by total information available. Scaled expected information gain associated with card x can then be defined as

$$SEI_g\ (x) = EI_g(x)/\sum\nolimits_{xi\ \ [p,\ not\ p,\ q,\ not\ q]} EI_g(x_i).$$

Using Hattori's (1999, 2002) 'selection tendency function' (STF) the probability that any particular card *should* be turned over is

$$P(T_x) = 1/1 + e^{2.37\text{-}9.06\ SEIg\ (x)}.$$

The STF maps scaled expected information gain on to the predicted probability that a particular card will be selected. This particular STF (Hatori, ibid) has also been used to map outputs of neural networks on to probabilities of responses. The parameters, 2.37 and 9.06 are Hatori's estimates based on past data from the selection task.

> Suppose that the hypothesis under test is ‘if a saucepan falls from the kitchen shelf (p) it makes a clanging noise (q).’ This rule, like the vast majority of everyday rules, conforms to the rarity assumption—saucepans fall quite rarely…and clangs are heard quite rarely….The four cards in the selection task can be seen as analogous to the following four scenarios. Suppose I am in the kitchen, and see the saucepan beginning to fall (p card); should I bother to take off my headphones and listen for a clang (i.e. should I turn the p card?)? Intuitively, it seems that I should, because, whether there is a clang or not, I will learn something useful concerning the rule (if there is no clang, the rule is falsified; if there is a clang, then my estimate of the probability that the rule is true increases). Suppose on the other hand that I am next door and I hear a clang (q card); should I bother to come into the kitchen to see whether the saucepan has fallen (should I turn the q card?)? Intuitively, this is also worth doing—if the saucepan has not fallen then I have learned nothing (something else caused the clang); but if the saucepan has fallen, then this has strongly confirms the rule….Now consider the analogue of the turning of the not-q card: I am next door and I hear no clang. This time should I bother to come into the kitchen to see whether the saucepan has fallen…?[89]

The answer to this question is obviously no. This RA approach to explaining the results of the Wason task seems successful to the extent that it can provide a rational explanation for the observed behavior of subjects. More generally, RA attempts to construct such models based on formal decision theory to explain a host of human reasoning behaviors. However, I have two reservations about this approach.

First, though the intuitive example seems right, I am not sure that the intuitive example *is* analogous to the elements in the Wason task. Though both examples are of rare events, falling sauce pans and clangs are closely associated in everyday situations with which people are familiar. Cards marked ‘D’, ‘3’, ‘K’, and ‘7’ are not familiar associations. We expect that when a saucepan falls, it will make a noise. We expect this because we have had experience with these

[89] Ibid., pp. 158-159.

kinds of events. When something falls, we expect it to make a sound. I think that this sort of expectation is part of our background information, and is taken for granted, because our broad experience supports the supposition that usually medium sized bodies that fall to the floor make a noise. But whereas Chater and Oaksford's analysis fits the everyday case of falling sauce pans, it does not seem to fit the highly contrived experimental conditions of the Wason task. We do not closely associate, Ds, 3s, Ks and 7s, and have no preconceived idea as to whether a rule for such associations may or may not hold.[90]

Second, and more importantly I think, this approach runs the risk of licensing bad reasoning behavior wholesale in less mundane cases. In the above cases not much is at stake when the card marked '3' is chosen, or when the clanging sound occurs and one checks for a falling saucepan. Not a lot seems to hang on whether we agree to treat these examples as falling under some kind of inductive reasoning strategy. When more is at stake, though, it seems that this model is highly suspect. Suppose that the stakes are raised in a different kind of

[90] In note 60, notice that when P(p) and P(q) are small, the information gain model predicts that $P(T_q) > P(T_{\sim q})$. This is because when P(p) and P(q) are small $EI_g(p)$ and $EI_g(q)$ become large. This is because these two cards, p and q, are more informative as to which hypothesis, M_D or M_I is true relative to the other cards. Oaksford and Chater argue (2007, p. 174) that the fact that these probabilities should be low is consistent with the way natural language works to carve up categories in the world. For example, 'thing' refers to objects and non-objects alike, so the scope of its reference is much broader than the scope of the reference of terms like 'desk'. The larger the scope of reference, the higher the probability that the term will refer to an object or non-object in the world and the smaller the scope of reference, the lower the probability of the term referring. This is the rarity assumption. Perhaps specifying that "D cards have 3s on the back" as part of the artificial set-up of the experiment would make the artificial set-up more like a non-contrived example (like falling sauce pans and sounds). This would make the rule part of our background information. However, I do not think this ameliorates the situation for Oaksford and Chater. Recall that their characterization of the selection task is that participants want to know which hypothesis, M_D or $M_{I,}$ accurately describes the relationship between the numbers and letters on the cards, and their task is to choose the data that will give them the most information in order to make this choice. While manipulating the artificial set-up of the selection task would serve to make it more like a natural set-up, it would not explain why choosing the specified cards is rational.

setup, but where a lot hangs on whether the analogous rule is true: agents who still "picked the wrong cards" with regard to fully testing the rule would suffer for it greatly in these circumstances. *Rational analysis* allows for this kind of change in situation because it requires that the environment be specified for each kind of case; only the results expected in a specific sort of environment are checked against experimental results for that type of environment.

Specifying the environment in an appropriate way may prove difficult in some cases. Furthermore, this raises the issue of how closely goals must be connected to an environment. Are specified goals so environment-specific that when the environment changes the goals must always change as well, or may goals remain the same across different environments? Here are two examples where intuitively it seems that the goals must remain the same across different environments.

In case 1 suppose the environment is a medical setting in which a medical professional in a developed country makes a diagnosis of HIV for a patient who tests positive after taking an HIV test. Suppose that the false positive rate for this test is well known and is around 5%. The judgment that the patient has HIV is too hasty when solely based on the results of a single test with this kind of false positive rate. By calling this judgment 'hasty' we are making a negative evaluation about the judgment itself. This negative evaluation about the judgment on the part of the medical professional seems right. This is because in an ordinary medical setting like this one we expect medical professionals to pay attention to the influence of false positive rates; that is just part of normal rationality.[91] A positive result on such a test calls for the patient to be retested with a more rigorous (more expensive) test that has a much lower false positive rate.

In case 2 suppose that the situation is similar to case 1, but is one in which the more rigorous test with the much lower false positive rate is not available. But suppose that anti-HIV drugs are relatively cheap and highly effective, and have few side effects. Then it might be rational to diagnose the patient as having HIV, even though the probability of HIV is still low after the positive test result. Better to misdiagnose and treat patients who do not have the disease than take the chance of not treating someone who has HIV.

In the two examples above, clearly which behavior counts as optimal differs. But both cases are within the context of scientific practice, where the aim is to be objective. Objectivity in this case is considered to be a normative good. The two cases show that rationalizing behavior in terms of a formal model is not completely satisfactory on its own. In the next chapter I propose to incorporate an additional normative component into the context of rational decision, that of intellectual virtue. I will try to show that it has an important role to play in assessments of whether an action or a decision is rational.

[91] If the base rate for HIV is 1/1000 and the false positive rate is 5%, the probability that a patient has HIV when his test is positive is less than .02.

CHAPTER 4

WHY VIRTUE EPISTEMOLOGY?

Introduction

Virtue epistemology (VE) consists of a broad family of theories with a common feature: what is central to knowledge and justification is the intellectual (and perhaps moral) character of the agent. One faction within this family emphasizes the truth-conducive reliability of intellectual character. Another faction emphasizes the responsibility that an agent has when assenting to belief, another aspect of intellectual character. What is common to both of these conceptions of VE is that the intellectual character of the agent is the focal point of analysis. Knowledge and justification derive from properties of persons as belief-acquiring agents. VE's shift in focus stands in contrast to the traditional conception of knowledge. The traditional conception holds that knowledge is merely a species of belief that is distinguishable from other, non-knowledge instances of belief. The mark of distinction between a knowledge belief and a non-knowledge belief is solely the properties that accrue to the belief. On this view knowledge is distinguished from other kinds of beliefs by identifying properties of particular beliefs, collections of beliefs or processes by which beliefs are acquired.

The aim of this chapter is to motivate VE by discussing three major issues that plagued 20th century epistemology—elucidating the structure of justification, elucidating the nature of justification, and the general insufficiency of accounts of knowledge as justified true belief. In shifting the focus of analysis from properties of the beliefs to the belief-acquiring properties of agents that make up epistemic (or, intellectual) character, VE appears to offer an alternative account of the nature of knowledge that fares better than alternatives.

This chapter proceeds as follows: Section 1 presents the foundationalism/coherentism debate in epistemology regarding the structure of justification, Section 2 presents the internalist/externalist debate regarding the nature of justification, Section 3 presents the central problem of defining knowledge in terms of justified true belief – i.e. the so-called *Gettier problems*, and in Section 4 I introduce and discuss some preliminary considerations regarding VE, and I argue that employing the strategy (adapted from virtue ethics) of drawing on the stable dispositions of the agent in belief-acquisition, rather than focusing on the properties of beliefs and their connections to other beliefs, provides an important turn in the conception of *knowledge*. In the next chapter I will develop my own conception of VE.

Section 1—The Structure of Justification

In this section I will elucidate the central issues raised in recent philosophical debates over the structure of epistemic justification. The issue of the structure of justification goes back at least to the Pyrrhonian problematic. The Pyrrhonian problematic can be stated this way: if there is no foundation upon which our justified beliefs rest, then an infinite regress of justification looms.

One way to think about the Pyrrhonian problematic is in terms of Alston's (1992) distinction between mediately and immediately justified beliefs. A mediately justified belief is a belief that relies on another belief for its justification. An immediately justified belief is a belief that does not.[92] The Pyrrhonist problem can be stated this way: if what justifies a belief B1 is itself a belief B2, then B2 must also be justified in order to justify B1. Another belief, B3, must justify B2, so B3 must itself be justified by some belief B4, and the regress continues unless there is a stopping point, a foundational belief that is immediately justified. Alternatively, the chain of justification may loop back into itself, circularly. The reason this is problematic is that it seems to lead to skepticism. Jonathan Kvanvig (2007) characterizes the skeptical argument arising from the Pyrrhonian problematic this way:

1. No belief is justified unless its chain of reasons is infinitely long, stops, or goes in a circle.
2. An infinitely long chain of reasons involves a vicious regress of reasons that cannot justify any belief.
3. Any stopping point to terminate the chain of reasons is arbitrary, leaving every subsequent link in the chain depending on a beginning point that cannot justify its successor link, ultimately leaving one with no justification at all.
4. Circular arguments cannot justify anything, leaving a chain of reasons that goes in a circle incapable of justifying any belief.[93]

[92] William Alston, “Foundationalism,” in *A Companion to Epistemology*, Jonathan Dancy and Ernest Sosa, (eds.), Blackwell: Cambridge, MA (1992), p. 382.

[93] Kvanvig, Jonathan, "Coherentist Theories of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/justep-coherence/>.

As Kvanvig notes, two main views regarding the structure of justification have emerged to address this problem, coherentism and foundationalism. Both agree that skepticism is false, but differ on how to solve the problem.

Foundationalism is the view that the structure of justification is ultimately anchored to non-inferential beliefs; so what ends the regress is a belief that is immediately justified. An immediately justified belief is a belief whose source of justification is non-inferential in nature. An immediately justified belief may be based on experience, or it may perhaps be self-justified.[94] For example, my belief that 2 + 2 = 4 is obvious in a self-evident way to me. So if self-evidence is a legitimate source of justification, then this belief is justified. Furthermore, suppose it appears to me now as though a tree is before me. If immediate experience is a source of justification, then the belief that a tree is before me is justified by how things appear to me now.

One thing to note: on Kvanvig's characterization of the skeptical argument, the foundationalist accepts premise 1 above—that the chain of reasons is infinitely long, stops, or loops back into a circle. The foundationalist goes for the thesis that the chain of reasons stops at base beliefs that are ether self-justified or justified some other way, independent of other reasons (e.g. justified from experience).

Coherentism regarding the structure of justification is most simply described as the denial of foundationalism. Since I am only concerned with giving an overview of coherentism, I will only discuss the most popular version of

[94] Alston, ibid.

this view. On what Kvanvig calls the standard account of coherentism,[95] justification is holistic, rather than atomistic. That is, justification is not linear in the way that the statement of the Pyrrhonian problem assumes, but is the property a belief set as a whole. So, the coherentist would reject 1 from above—the chain of reasons neither stops (in the straightforward way that foundationalism claims), nor continues infinitely, nor loops back into a circle. This is because justification is a property of a system of beliefs, where, if a belief system is coherent, a belief within that system is justified by its membership within the coherent system. This holistic view stands in contrast with the atomistic view of justification that individual beliefs are justified by other individual beliefs.

Coherentists maintain that justification is a relation among propositions. Citing Neurath, Sosa comments on the metaphor that is commonly used to describe (in broad outline) the basic coherentist idea.

> "…our body of knowledge is a raft that floats free of any anchor or tie. Repairs must be made afloat, and though no part is untouchable, we must stand on some in order to replace or repair others….what justifies a belief is not that it be an infallible belief with an indubitable object, nor that it have been proved deductively on such a basis, but that it cohere with a comprehensive system of beliefs."[96]

The biggest challenge for the coherentist view is explaining the coherence relation itself. I will return to this issue below, after first discussing some of the main problems faced by foundationalism.

[95] Kvanvig, ibid.

[96] Ernest Sosa, "The Raft and the Pyramid" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 147.

One of the major problems with foundationalism involves establishing just what the foundational beliefs are supposed to be. The foundational belief must either be a cognitive state or a non-cognitive state of an agent. If it is a cognitive state, then the challenge for the foundationalist is to show how the immediate belief is tied to truth—to a way the world actually is. If it is a non-cognitive state, then the challenge for the foundationalist is to show how a non-cognitive state can serve as justifier for a belief set consisting of what are presumably cognitive states of an agent. The difficulty lies in the fact that cognitive mental states have propositional content, and non-cognitive mental states do not.

One useful way to understand the problem for foundationalists who hold that the foundational belief is a cognitive state starts with incorporating a distinction that Sosa uses. He points out that a propositional attitude is a special kind of mental state that incorporates a proposition. While we have other mental states, like headaches, the headache itself does not incorporate a propositional attitude, nor is it constitutive of a propositional attitude.[97] This distinction is useful because it points to the fact that propositions are truth bearers, which when incorporated by a mental state or constitutive of a mental state involve judgment. Fumerton (2010) characterizes the problem this way:

> It is crucial that the foundationalist discover a kind of *truth* that can be known without inference. But there can be no bearers of truth value without judgment and judgment involves the application of concepts. But to apply a concept is to make a judgment about class membership, and to a make a judgment about class membership always involves relating the

[97] Ibid.

> thing about which the judgment is made to the paradigm members of the class.[98]

If the foundational belief, the mental state, is a propositional attitude, then the proposition is a bearer of a truth value, but this involves applying a concept, and applying a concept involves a judgment. Judgment is inferential because it will at the very least involve concepts or beliefs about the past. But, since the foundational belief is supposed to stop the regress of inferential justification, then it must be arrived at non-inferentially. So, this objection goes, the foundational belief cannot be a cognitive state.

The objection to the version of foundationalism that holds that the basic belief is non-cognitive also draws on Sosa's distinction. To see the objection, first suppose a non-propositional mental state can serve as a foundation for justifying other beliefs. So, for example, consider the mental state of having a headache. This is a mental state that does not have propositional content, and so it is non-cognitive. Suppose that you form the belief that you have a headache. There seems to be no obvious way for a state such as having a headache, one without propositional content, to justify a state with propositional content, such as the belief that you have a headache. This is because, as Bonjour points out, the non-cognitive state involves no cognitive grasp of the state of affairs. Since your non-cognitive state does not involve your cognitive grasp of your particular state

[98] Richard Fumerton, "Foundationalist Theories of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/justep-foundational/>.

of affairs, it is difficult to see how your non-cognitive state can give you a reason to think that your belief that you have a headache is true.[99]

While these arguments against the cognitive-basis and non-cognitive-basis versions of foundationalism are insufficient to destroy the position entirely, I think that it is sufficient to cast doubt on foundationalism as an adequate theory of knowledge or justification. On the one hand, it is true that some of our beliefs are justified inferentially, and it might be useful to characterize the chain of inference for some beliefs in terms of a structure that rests upon some kind of foundational belief. For example, my belief that Lucy is a mammal is justified inferentially by other beliefs in my stock of beliefs. Two beliefs that support it directly and inferentially are: *all cats are mammals* and *Lucy is a cat*, and I might stop there if offering reasons for my belief, though neither of those beliefs is a foundational belief that could stop a regress of reasons for holding the belief that Lucy is a mammal. On the other hand, since neither of these beliefs (that all cats are mammals and Lucy is a cat) could stop a regress, I could go on offering reasons for those beliefs, and even though I might be able to eventually stop the regress of reasons by positing a state that is either self-evident or self-justifying, I would still have to establish whether or not this foundational state is cognitive or non-cognitive. So, while foundationalism regarding the structure of justification might be useful for some purposes (e.g. characterizing the structure of justification for some kinds of beliefs), it is doubtful to me that it captures the whole story about justification, and so it is doubtful to me that it is an adequate theory of knowledge.

[99] Laurence Bonjour, "Can Empirical Knowledge Have a Foundation?" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 117.

Since foundationalism seems inadequate, let us now return to coherentism as an alternative to foundationalism. There are three major problems with coherentism. The first problem for the view is determining exactly what coherence among beliefs is. The straightforward answer, logical consistency, is too weak. This description of coherence is too weak because mere logical consistency of belief does not imply that any of the beliefs are in fact true. It is entirely possible to have a consistent system of beliefs that are almost wholly false. As Kvanvig notes, "a good piece of fiction will display the virtue of coherence, but it is obviously unlikely to be true."[100] One reason this is such an important objection to coherentism is that a central value in epistemology, though, perhaps not the only value, is the goal of having a large body of non-trivial true beliefs, and very few non-trivial false beliefs.[101] A piece of fiction may exhibit all of the features lauded by Bonjour and still be false with regard to almost all non-trivial claims. Although truth may not be the only epistemic value, it is surely among the most important, and an adequate epistemology should surely tie justified or "correctly held" belief to truth in a way in which justification or "correct holding" of the belief tends towards truth.[102]

[100] Kvanvig, op. cit.

[101] The qualification about trivial truths is needed because if we only count numbers of truths, we may, for example, enumerate truths stating which letters of the alphabet fail to occur at each small region of the pages in the books in the Library of Congress. That's a huge number of truths, but a system of beliefs that was only correct about these truths, and got substantive, useful claims wrong, would hardly figure as a significant body of knowledge.

[102] For an insightful discussion of value in epistemology, see Wayne Riggs', "The Real Value of Knowing that P" in *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* Vol. 107, No. 1 (Jan., 2002), pp. 87-108. The problem for coherentism of linking a coherent belief set to the world brings Bonjour to abandon coherentism in later publications. Cf. "Foundationalism and the External World," in *Noûs,* Vol. 33, Supplement: *Philosophical*

A second problem, closely related to the first problem, is that there is no guarantee of cognitive contact with reality for an agent with a coherent belief set.[103] This means that a collection of beliefs could be logically consistent, all of them in fact true, but lack a property that is central to knowledge—the appropriate connection between the believer and the facts of reality. This connection with reality, which I can only call 'cognitive contact' , is what distinguishes knowledge from ordinary opinion or lucky guess. Ordinary opinion and lucky guesses may be consistent with other beliefs in a belief set, and each may be true, but fall short of knowledge because each lacks that cognitive with reality.

The third problem that coherentism faces is explaining how the coherence relation among beliefs confers justification on individual beliefs. One answer is from Bonjour (1985) who characterizes coherence as being "a matter of how well a body of beliefs "hangs together": how well its component beliefs fit together, agree or dovetail with each other, so as to produce an organized, tightly structured system of beliefs, rather than either a helter-skelter collection or set of conflicting subsystems."[104] Bonjour spells out this idea more fully in terms of the logical and probabilistic consistency of beliefs in the belief set, the inferability of the beliefs in the set from others in the set, the explanatory relations among the beliefs in the set, the relatedness and connectedness of the beliefs in the set, the lack of

Perspectives, 13, Epistemology (1999), pp. 229-249 for Bonjour's subsequent defense of foundationalism.

[103] Admittedly, this is also a problem for foundationalism. As a result, both foundationalism and coherentism require adequate theories of perception in order to address this problem.

[104] Laurence Bonjour, *The Structure of Empirical Knowledge*, Harvard University Press: Cambridge, MA (1985), p. 93.

anomalous beliefs in the set, and the degree to which conceptual changes can be accommodated within the set.[105]

Bonjour’s account makes a kind of intuitive sense. However, as I see it, the problem for coherentism in general (and not merely this version of it) is to articulate how these factors are supposed to come together, each playing an appropriate contributing role to the justification of individual beliefs. For example, consider the lottery paradox.[106] This paradox centers around the issue of rational acceptance. If a belief is justified, then, presumably, it is rational to accept it. Suppose we have a fair lottery with 1000 tickets. The belief that exactly one ticket will win is justified because it is justifiably believed to be a fair lottery. However, one might also justifiably believe of each individual ticket that it will not win—that ticket 1 will not win, that ticket 2 will not win,…, that ticket 1000 will not win—because the chance of each individual ticket winning is so low (1/1000). The conjunction of each of these beliefs about each individual ticket is logically equivalent to the belief that no ticket will win the lottery, which contradicts the justified belief that exactly one ticket will win. One of the features of Bonjour’s coherentist view is that coherence is a function of the degree to which beliefs are logically consistent. Obviously, believing that exactly one ticket will win and also that no ticket will win is inconsistent, and believing this inconsistency is incoherent (some would say “irrational”), though there seems to

[105] Ibid., pp. 95-101.

[106] Bonjour addresses this paradox in the context of it as a defense of externalist foundationalism which ultimately fails to support the externalist foundationalist thesis (pp. 53-57). For an informative discussion of the lottery paradox in general, cf. John Hawthorne’s *Knowledge and Lotteries*, Clarendon Press: New York NY (2004), pp. 1-50. Hawthorne gives multiple versions of the paradox in order to show the depth of the problem. My presentation above is a simplified version of the paradox, designed to exhibit a major problem for coherentism.

be sufficient justification for believing that exactly one ticket will win and for believing of each ticket that it will not win. There is an apparent anomaly here, and it is not clear how this anomaly is to be resolved. If it is incoherent (or irrational) to believe a contradiction and you are justified in believing each component of the contradiction, how is the coherence-as-knowledge view to avoid incoherent belief? Some have argued that an agent need not believe the conjunction of her individual beliefs, and indeed in cases of this sort should not do so. But this seems a very counter-intuitive move to most epistemologists.

In this section I have merely given a brief sketch of the historical debate between coherentists and foundationalists and described some of the main problems with each. I do not claim that either view has been refuted. But I hope to have shown that at the very least both views are problematic – both have problems that remain unresolved. Both seem to appeal to common intuitions, but these intuitions often clash. In the next section I will describe the internalist/externalist debate over the conditions for justification. The debate between these two positions also remains unresolved.

Section 2—The Conditions for Justification

Another dimension of the nature of justification is the internalism/externalism debate. Internalism about justification is a broad thesis about the basis for justification. There are several distinct versions of internalism, and several distinct internalist theses.[107] For each version of internalism there are

[107] Cf. George Pappas "Internalist vs. Externalist Conceptions of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2008/entries/justep-intext/ for ways to distinguish between various forms of internalism.

corresponding externalist responses that deny a corresponding central internalist claim. For the sake of tractability I will be concerned only with the nature of justification. By this I mean what Sosa, et al. refer to as the ontological aspect of internalism.[108] The ontological version of internalism only requires that justifiers for a belief be part of the mental contents of the agent.[109] Externalism, on this view, denies this claim. An externalist view is one where the justifier need not be part of the mental contents of the agent. In this section I will provide a broad characterization of ontological internalism (from here on out that's what I will mean by 'justification internalism' or simply 'internalism') and externalism about justification and their associated problems.

Perhaps the best known version of justification internalism is Feldman and Conee's view, which they call *mentalism*. According to Feldman and Conee, "A mentalist theory may assert that justification is determined entirely by occurrent mental factors or by dispositional ones as well. As long as the things that are said to contribute to justification are in the person's mind, the view qualifies as a version of mentalism."[110] On this view, the only things that serve as justifiers are occurrent mental contents or dispositional factors (such as the disposition to form true perceptual beliefs). The defense that Feldman and Conee provide is

[108]Ernest Sosa, "Introduction to Part V" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 306.

[109] Indeed, Feldman and Conee further divide internalism into *access internalism* and *mentalism*, where the refined version of *mentalism* is the view that beliefs are only justified by things internal to an agent's mental life. Cf. "Internalism Defended" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 408.

[110] Richard Feldman and Earl Conee, "Internalism Defended" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 408.

straightforward. In any case of justification, a mental state is responsible for the justification, and in any case where a belief is not justified, a mental state or lack thereof is responsible. Feldman and Conee make their point by means of examples. Here are two.

> Example 1: Bob and Ray are sitting in an air-conditioned hotel lobby reading yesterday's newspaper. Each has read that it will be very warm today and, on that basis, each believes that it is very warm today. Then Bob goes outside and feels the heat. Both continue to believe that it is very warm today. But at this point Bob's belief is better justified.[111]
>
> Example 2: Hilary is a brain in a vat who has been abducted recently from a fully embodied life in an ordinary environment. He is being stimulated so that it seems to him as though his normal life has continued. Hilary believes that he ate oatmeal for breakfast yesterday. His memorial basis for his breakfast belief is artificial. It has been induced by his "envatters."[112]

There are two versions of details relevant to this second example.

> 2a) Hilary's recollection is very faint and lacking in detail. The meal seems incongruous to him in that it strikes him as a distasteful breakfast and he has no idea why he would have eaten it.
>
> 2b) Hilary's recollection seems to him to be an ordinary vivid memory of a typical breakfast for him.[113]

In Example 1 the relevant epistemic difference between Bob and Ray lies in something internal to each agent. In this case Bob is better justified than Ray in the belief that it is very warm today. Assuming that both Bob and Ray obtained the weather report from a reliable source, they are both initially justified in their belief that it is very warm today; and given the description of the

[111] Ibid., p. 409.

[112] Ibid., p. 410.

[113] Ibid.

circumstances, I think it is safe to say that they are equally justified up to the point before Bob goes outside. However, after going outside, Bob is better justified in his belief that it is very warm today because he went outside and experienced the temperature – he "'internalized' the actual temperature"[114] with his experience of the warmth of the day. When he went outside he gathered additional evidence for his belief that it was warm outside.

In Example 2, the relevant epistemic difference between versions 2a and 2b is just an internal difference in introspective mental state for Hillary. In (2a) Hillary's belief is not what Feldman and Conee call "well justified."[115] The means by which the belief is formed, recollection, is weak. The memory itself is one that Hilary might well come to question or doubt. It is incongruous with the rest of his beliefs. In (2b), however, the belief is well justified. The means by which the belief comes about is of the ordinary strength possessed by many other beliefs (including those that might well also count as knowledge), and the belief is congruent with other beliefs that Hilary would have.

Feldman and Conee put the point like this. "It is reasonable to generalize from these examples to the conclusion that every variety of change that brings about or enhances justification either internalizes an external fact or makes a purely internal difference. It appears that there is no need to appeal to anything *extramental* to explain any justificatory difference."[116] On this view the only factors that serve as justifiers are factors that are internal to an epistemic agent,

[114] Ibid., p 409.

[115] Ibid., p. 410.

[116] Ibid.

and specifically they are only the mental contents of the agent. I will address problems with this view below. But before doing so I first want to characterize the contrasting view called *externalism*.

Externalism denies the internalist claim that there is no need to appeal to anything *extramental* to explain justificatory differences. In the first example, an externalist might well claim that the difference in justification between Bob and Ray is in fact *extramental*. Bob's evidence is in part constituted by conditions external to his mental state. If his sensory apparatus for detecting temperature were not functioning properly (e.g. he was suddenly hit by a high fever that he wasn't aware of) when he went outside to the check the weather conditions, then his sensation of the heat would fail to be justificatory (despite his internal mental contents), or so an externalist might argue.

To illustrate externalism's denial of the internalist thesis, consider one early version of the externalist view called *reliabilism*. Alvin Goldman's *process reliabilism* is a version of externalism that asserts that factors external to the agent's mental content are crucial to justification. It is the view that the justificational status of a belief is dependent on the reliability of the causal process or processes involved in forming the belief. Reliability is based on the tendency of a process to produce true beliefs rather than false beliefs.[117] Goldman admits that he has to remain vague on the probabilistic threshold associated with reliability, on what counts as a "tendency", and on the precise nature of what

[117] Alvin I. Goldman, "What is Justified Belief?" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 338.

counts as a process. However, he does not find this to be problematic because our ordinary notion of justifiedness is equally vague.[118]

Goldman's process reliabilism takes Bob's belief-forming processes to be the key. The belief forming process involved in his belief that it is very warm outside when he steps outside is reliable *provided that* most of the time, under similar circumstances, Bob tends to form true beliefs about environmental conditions in which he finds himself. Reliability of the process confers justification, not the mental contents of the agent. Because the process itself is extramental (i.e. not part of the mental contents of the agent), this is an externalist position. So, Bob may have a highly reliable belief-forming process involving environmental conditions, but not be occurrently or dispositionally aware of the process.

Does the fact that Bob checks the temperature by going outside make Bob more justified than Ray, who relied entirely on the weather report? Maybe. Although the newspaper report about the weather that day may be reliable (assume that this is a highly reputable publication), Bob's belief-forming process involving his experience of the weather, combined with his belief forming-process involving his reading the reputable newspaper, may add to the justificatory status of the belief. Assuming that Bob's belief-forming process about environmental conditions is more reliable than his belief-forming process involved in reading the weather report, Bob has better justification for his belief about weather conditions than he would if he had only read the weather report. This is because Bob's environmental condition belief-forming process tends to get

[118] Ibid., p. 339.

Bob more true beliefs than Bob's belief-forming processes involving reading reputable newspapers.. If Bob's belief-forming process about environmental conditions is not more reliable than, but only equally reliable to, his newspaper-trusting process, then perhaps the combination of the two processes adds to the justificatory status of the belief. If Bob's belief-forming process about environmental conditions were the less reliable of the two, then it would not add to the justificatory status of the belief.

Hilary's case, though, is a bit more tricky because the thought experiment starts with an archetypal skeptical situation. In this situation, Hilary could never have the kind of justification that leads to knowledge, according to process reliabilism. Perhaps he has something "justification-like" that would count as real justification were he not a brain in a vat. But envatted as he is, his belief-forming process is unreliable. Goldman takes memory to be a reliable belief-forming process. So, if we change the example to avoid the skeptic (i.e. Hilary is not a brain in a vat, but has the breakfast belief as a result of his memory of his breakfast), and if memory is a belief-forming process that is sufficiently reliable to confer justification, then Hilary would be justified in his breakfast belief, and he would be justified in the same kind of way that anyone is justified when they have knowledge. Notice that the internalist would have to say that envatted Hilary is justified in his belief (although, unfortunate in his situation), whereas the externalist can bring the external situation in which Hilary is embedded into the justificatory account, and thus deny that Hilary is justified. Envatted brains fed false sensory information do not obtain their beliefs via a reliable process.

Since the dominant forms of externalism about justification and knowledge are forms of reliabilism, I want to address a common objection to reliabilism—the so-called *generality problem*. The generality problem is a longstanding problem raised against process reliabilism. The problem is how to specify the appropriate *type* of process involved in a token instance of reliability (qua justification). Each individual case of belief formation (i.e. each token of belief formation) falls under a wide variety of formation-process-*types*, some of them reliable (to various degrees) and some of them unreliable. For example, consider the belief that I have as I write this, *I see my laptop*. On the process reliabilist view, if it is a justified belief (and I think it is), I form this belief as a result of a token causal process that is reliable. Suppose that the token causal process is the following.[119] Light reflects off of my laptop into my eyes, which causes optic neural events which further cause other neural events that ultimately lead to the formation of the belief that I see my laptop. The conjunction of these events (e.g. light reflecting off my eyes, the resulting optic neural events, the resulting neural event involved with applying concepts, etc.) is a token instance of the causal process involved in the formation of my belief. Just because this particular token instance resulted in my true belief, I cannot rightly say that this instance is reliable. Reliability, originally conceived by Goodman and I think understood generally, is a tendency. In terms of belief formation, the tendency is to produce true beliefs rather than false beliefs. So, a process *type* has to be reliable, has to have the tendency to produce true beliefs, rather than a process

[119] This is an adaptation of Feldman and Conee's original example in "The Generality Problem for Reliabilism" in *Philosophical Studies* Vol. 89: 1–29 (1998).

token. This token instance of the causal process can by typed in various ways. For example, this token can by typed as a process of involving the visual faculty, as a process the prompts my particular beliefs about my computer, or as a process that occurs when I am sitting on my couch. The number of types this token process can fall under is limitless, and the reliability of each token instance depends on the type under which it falls. Different types of causal process have different degrees of reliability. The issue is to decide first which *type* is supposed to apply in a token instance and second to what degree must the type be reliable for the process to confer justification on the belief.[120] Indeed, I think the generality problem is a problem for any version of externalism.[121] This is because all versions of externalism trade on reliability in some way, and the specification of reliability is always relative to a *type* of situation. In other words, for each token instance of reliability, whether it is in the form of belief formation processes or some other form, the conditions for the situation or process to count as *reliable* need to be specified. The appropriate types will often hang on things like the level of detail that is relevant to the epistemic situation involved. However, the most significant issue that the generality problem raises is this: there is no reason

[120] Richard Feldman and Earl Conee, "The Generality Problem for Reliabilism" in *Philosophical Studies* Vol. 89: 1–29 (1998)., pp. 2-3.

[121] The version explicated in this chapter is not the only one. There are others such as the truth-tracking view. The truth-tracking view originated with Nozick (1981). The idea is that for a belief to be justified it must be the case that the belief was formed in such a way that were it not true, the agent would not have come to believe it. For an interesting update on Nozick's truth-tracking view that seems to avoid the generality problem see Chapter 3 of *Tracking Truth: Knowledge, Evidence, and Science,* by Sherrilyn Roush, New York: Clarendon Press (2005). Here Roush avoids the generality problem by specifying two rules to fix the reference class applicable to a subject's belief-forming situation when the subject comes to believe a proposition.

to suppose that there is a fact of the matter about which specific process-type is the one that determines justification in any given case.

Furthermore, the generality problem reaches beyond externalist reliabilist epistemologies. For example, any *virtue theory* of epistemology that emphasizes the reliability of the agent will have to address it. (I will address virtue theories a bit later.) However, I do not think that the generality problem is unsolvable in general. But how it may be solved or resolved will depend on the details of the individual epistemic theories where it arises.

The main problem for internalism, as I see it, is how an internalist accounts for the "internalization" of external facts. One prominent way to do this is to commit to a foundationalist structure for knowledge where at the foundation is a base belief formed from sense experience. However, if this move is made, then the internalist foundationalist has to resolve the foundationalist problems discussed earlier. Is a base belief a cognitive state or a non-cognitive state? If it is a non-cognitive state, then he has to account for how a non-cognitive state can serve as a justifier for cognitive states. Alternatively, if the internalist takes a base belief to be a cognitive state, then he needs to give an account of justification that works independently of the external environment in which cognition takes place. If the justificatory conditions are purely mental contents, independent of the external world, then a brain in a vat could be justified in his beliefs, even though none of them are true. We typically think that truth and justification are intimately linked, so much so that when we have adequate justification (e.g. in the form of evidence) we think that the belief is likely to be true. We take the link between truth and justification so seriously that when the link is broken we do not count the belief as knowledge. The issue is, when the link is so broken that the justification of a belief no longer even makes the truth of the belief at all probable, should *that* still count as *epistemic justification*.

Section 3—The Problem of Defining Knowledge

Before moving on to a general discussion of virtue epistemology, I want to discuss briefly the problem of defining knowledge. While this problem did not originate the 20th century, it came to the center of discussion with the famous Gettier paper.[122] The crux of the Gettier problem is this. Justified true belief is not sufficient for knowledge. It is uncontroversial that one can have justified false beliefs. Because of this, standard definitions that rely on justification as the component that turns true belief into knowledge fail. For example, if we take the definition of knowledge to be:

S knows that P iff: (a) P is true
(b) S believes P
and (c) S is justified in believing that P

then we can construct an example where all three conditions are met, but S fails to know.[123]

One important way to understand Gettier cases is Zagzebski's 1994 analysis. In that paper, she points out that the historical response to this problem

[122] Edmund Gettier, "Is Justified True Belief Knowledge?" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), pp. 192-193.

[123] Cf. ibid., for specific cases.

in the time since Gettier's publication is to either redefine justification so that it is sufficient for knowledge or to add an extra component to the standard definition so that the resulting definition (usually justified true belief plus something else) is sufficient for knowledge.[124] Though she is not the first person to acknowledge that these are two common responses, I think that hers is an important way to acknowledge the problem. Since the publication of the problem, the standard response has been to change the definition of knowledge by focusing on properties of beliefs (e.g. redefine justification such that as a property of a belief it turns true belief into knowledge or add another property to justified true belief such that this property turns justified true belief into knowledge) rather than properties of the agent who holds the belief. However, none of these responses are immune to Gettier-type cases. As Zagzebski argues, as long as truth is independent of justification (or some other property attributable to a belief such that this property is strong enough to turn true belief into knowledge) a Gettier case can be derived to undermine the theory of knowledge.[125]

Another important point that she makes is that Gettier problems are virtually unavoidable in any definition of knowledge as "true belief plus something else"[126] because of accidental features of that are inherent in Gettier cases. The accidental features are instances of luck where one instance of luck cancels out another instance of luck such that, though the agent has a justified true

[124] Linda Zagzebski, "The Inescapability of Gettier Problems" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008). p. 207.

[125] Ibid., p. 209.

[126] Ibid., p. 207.

belief, she does not have knowledge.[127] For example, consider the following case. Jones thinks that Smith is a cyclist who does not wear a helmet. Jones believes this on the basis of good evidence, such as having seen someone who she believes is Smith on a bicycle, riding in the rain, without a helmet. She forms the belief that Smith is a cyclist who does not wear a helmet, and as a result, resolves to tell Smith that he should wear his helmet, especially in the rain. Later in the day, she meets Smith and learns that Smith was in fact riding to class in the rain without his helmet, and that he regularly rides without his helmet. However, unbeknownst to her the cyclist that she saw earlier in the day was not Smith, but his twin brother, Brown, who happened to be riding on the path that intersected Smith's walk earlier in the day. Now her belief that Smith is a cyclist who does not wear a helmet is true and justified, but it is not knowledge. She is unlucky in the first instance because the evidence that she has to support the belief does not in fact support the particular belief that Jones is a cyclist, but lucky in the second instance because the belief just happens to be true. The bad luck involved in the first instance (her bad luck to have mistaken Jones for Brown) is cancelled out by her good luck in the second instance (her good luck to have formed a true belief, albeit accidentally).

I think that Zagzebski's analysis of Gettier problems is an important way to understand them. Her analysis points to the problem of focusing on properties of beliefs alone for accounts of knowledge rather than properties of agents. Properties like truth, justification or warrant, when attributed to beliefs, are not enough to account for what we take to be a robust sense of the term "knowledge."

[127] Ibid., p. 208.

True, justified, or warranted beliefs are not always knowledge even if a particular belief has all of these properties. For example, consider the Gettier case above, but suppose it is not a rainy day. Suppose that the visual conditions are perfect and that Smith and Brown are identical twins so alike that only their mother can tell them apart. Suppose further that Jones has even heard from reliable colleagues that Smith is a cyclist. In this modified version, Jones' belief is true, apt, justified, and warranted, but Smith still fails to know.

However, in a more robust way that pertains to knowledge attributions, I think, we mean to evaluate the agent who holds the belief. For example, when we evaluate a belief as unwarranted, we make an evaluation of not only the belief, but also the agent who holds the belief. We might even say of the agent that *he* is unwarranted, and in saying this we mean that he does not have the *right* to hold this belief for various reasons. We might hold him responsible in some sense for the lack of warrant because he is intellectually lazy (as with a lucky guess) or intellectually indiscriminate (as with beliefs that are held without appropriate scrutiny of the truth). We are obligated in some sense to have warrant for at least some of our beliefs, if not all.

Consider a case where someone forms a belief as a result of wishful thinking, the belief *I will win the next Powerball payout*. Suppose that this person knows the chances of winning on a single play and buys ten tickets with their last ten dollars. They form the belief, *I will win the next Powerball payout,* solely on the basis of their wishful thinking. Clearly, this belief, *I will win the next Powerball payout* is unwarranted, and anyone would consider the belief a *bad* kind of belief for someone to have. What makes it a *bad* kind of belief is *how it was formed.* It was formed in a way (wishful thinking) that usually does not lead to true beliefs. Moreover, and more importantly, when we evaluate the belief, we also evaluate the person who holds the belief. We would think that this person who holds this unwarranted belief is not entitled to hold the belief because of the way that she formed the belief.

This kind of evaluation applies to persons holding warranted beliefs as well—Gettier cases seem to show this. In a Gettier case understood in the way that Zagzebksi describes as, first, a case in which the truth of the belief and its justification are independent, and second, a case of good luck canceling out bad luck, though the belief has the properties of being true and being justified, the agent fails to know. Though this failure to know is not the fault of agent, the accidental features of the case take away any credit that we may give to the agent for holding the belief in the first place. Indeed, the accidental features of the case cause the agent to lose the credit that is associated with knowledge attributions. Credit is a kind of evaluation, and in cases of knowledge, it is a kind of positive evaluation. Though more will be said below about the notion of credit as it relates to knowledge, here I merely want to establish it as a term of evaluation in order to show that there is a link between evaluative property and knower.

This link between certain evaluative properties and knowers is similar to the link between certain evaluative properties and agents in moral action. A moral action is good at least in the sense that it brings about a good state of affairs, for example, but we also think that what constitutes part of the goodness of an act is the intention of the agent who performs the act. For example, suppose I am redeeming cash back points from a credit card, and I have the option of having the cash deposited to my checking account or donated to relief efforts in Japan. Suppose further that I fully intend to take the cash deposit to my checking account, but because the button for donating my rewards to the relief effort in Japan is so similar in visual presentation to the button for depositing the money into my checking account and they are spaced very closely together that I accidentally click on the button to donate to relief efforts. My accidental contribution is a good in the sense that it brings about a good state of affairs, but I cannot be evaluated as good because I accidentally donated the money.

A similar thing occurs in Gettier cases. A true belief can be justified by evidence, for example, but for a true belief to be justified sufficiently for knowledge, it cannot have the accidental features of Gettier cases. Recognizing the accidental features involved in Gettier cases points to the significance of the agent's role in knowledge attribution. For knowledge attributions, justifiedness, for example, is not merely a property of a belief. It must also be a property of an agent or at the very least causally related to the agent in some way because agents certainly play a role in accounts of knowledge. Virtually every theory of knowledge acknowledges this.

Furthermore, for some instances of knowledge attribution, the intention of an agent to know does not seem to matter. By 'intention to know' I mean the deliberative, internal, intellectual action that an agent may take in order to get to the truth. This deliberative action could be as simple as adjusting one's focus in order to verify what is seen or as complex as critical reflection. For example, in cases of simple perceptual knowledge (e.g. I know that I see my laptop in front of me and I am not entertaining any skeptical hypothesis), intentions do seem not enter in. (I say 'do not seem to' because in this instance I do not require verification of any kind to just know that my laptop is in front of me, and critical reflection is certainly not required.) Human beings have evolved to be able to acquire this kind of knowledge easily enough in most cases with minimal effort, deliberative or otherwise. However, for other instances of knowledge attribution, the deliberative action of an agent does seem to matter, as in cases where knowledge is acquired through intellectual effort in a non-accidental way. For example, I know that the capitol of Brazil is Brasilia because when I searched for the answer to the question, *What is the capitol of Brazil?* years ago I intended to get the answer, and being the kind of intellectual agent that I am, I consulted a reliable, trustworthy source for the answer.

Section 4—Virtue Epistemology

One response to the problems of the structure of justification, the nature of justification, and defining knowledge is to shift the focus of analysis from properties of beliefs to properties of agents.[128] Rather than analyze the notions of justification and knowledge solely in terms of properties of beliefs in isolation as if they stand independent from an agent, the approach from virtue epistemology is to look to properties of agents which somehow confer justification (or some other

[128] John Greco and John Turri, "Virtue Epistemology", *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/epistemology-virtue/>.

epistemically important attribute) onto beliefs such that these beliefs are knowledge-worthy. Indeed, Sosa's 1980 paper, "The Raft and the Pyramid" is widely acknowledged to be the inspiration for the resurgence of an interest in virtue epistemology. Here Sosa remarks on the ethical analogue to epistemology. What Sosa calls "reliabilism" in ethics is the view that "action [for an agent] is the result of certain stable virtues, and there are no equally virtuous alternate *dispositions* that, given his cognitive limitations, he might have embodied with equal or better total consequences, and that would have led him to [another, more dire action] in the circumstances."[129] The analogue to epistemology begins with justification and the different types of justification. "The important move for our purpose is the stratification of justification. Primary justification attaches to virtues and other dispositions to act, through their greater contribution of value when compared with alternatives. Secondary justification attaches to particular acts in virtue of their source in virtues or other such justified dispositions."[130] For epistemology, "…primary justification would apply to *intellectual* virtues, to stable dispositions for belief acquisition, through their greater contribution to getting us to the truth."[131] Secondary justification works the same way as in the ethical analogue. A belief is virtuous in virtue of its source in virtues or stable dispositions.[132]

[129] Ernest Sosa, "The Raft and the Pyramid" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), p. 159.

[130] Ibid.

[131] Ibid.

[132] Ibid.

So, what is an intellectual virtue? There are multiple and varied understandings, but all virtue accounts hold that intellectual virtues are excellences of the intellect (though, these excellences are not necessarily truth conducive). Here is a brief list of the epistemic virtues that are typically cited in VE accounts:

> attentiveness, courage (of the intellectual variety – willingness to buck the consensus when appropriate), creativity, curiosity, discernment, discretion, foresight, honesty, humility (e.g. willingness to admit that one might be wrong), imaginativeness, objectivity (e.g. attentiveness to justification and/or warrant for belief), parsimony (good judgment in the application of Occam's razor), perceptiveness, prudence, practical wisdom, studiousness, tenacity, truthfulness, understanding.

As with all virtues, these intellectual attributes turn to vices when taken to extremes. For example, intellectual courage turns to a kind of rashness or plain contrariness if not moderated by discernment and humility; and tenacity turns to stubbornness (pigheadedness) when not properly kept in check by humility.

Typically, accounts of intellectual virtues include a success component, and, somewhat less commonly, a motivational component. The degree to which the success component or the motivational component is emphasized varies with the account, and indeed, the varying degree of emphasis is one way to divide up kinds of VE. On the issue of excellences of the intellect, accounts which emphasize the success component of a virtue take the excellence to be attributed more to the reliability of the agent. In other words, agent reliability in terms of getting to the truth in a way that is attributable to the agent is what makes the virtue an excellence. Accounts which emphasize the motivational component of the virtue over the reliability of the agent take virtues to be more of an excellence of the agent's cognitive character. In other words, the motivational component of the intellectual virtue is what makes it an intellectual excellence, more so than the reliability of the agent.[133] While reliability is a necessary condition for knowledge, so is motivation on this kind of view.

So, there are at least two kinds of accounts in VE. One is an account of virtue emphasizing the reliability of cognitive faculties. These accounts are standardly called virtue reliabilist theories because they are descendent from earlier externalist, reliabilist theories of knowledge and justification. On these views, the emphasis is on the reliability of the agent's character to get to the truth. For example, Greco's agent reliabilism, an early form of his virtue reliabilism, defines a belief with positive epistemic status (where positive epistemic status is that which turns true belief into knowledge) as one that results from stable and reliable dispositions that make up an agent's cognitive character.[134] So, the emphasis on this account is the reliability of the agent's cognitive disposition to form true beliefs in the "right way" where the "right way" is the way that one would form a true belief that is sufficient for knowledge. For example, one aspect

[133] For two very clear examples of the distinction between accounts of VE which emphasize one component of virtue over another see Linda Zagzebski's, *Virtues of the Mind*, Cambridge University Press: New York, NY (1996) and John Greco's *Achieving Knowledge: A Virtue Theoretic Account of Epistemic Normativity.* Cambridge University Press: New York, NY (2010).

[134] John Greco, *Putting Skeptics in Their Place: The Nature of Skeptical Arguments and Their Role in Philosophical Inquiry*, Cambridge University Press: New York, NY (2000), p. 177.

of a belief being formed in the “right way” is that it is not formed accidentally, as in a lucky guess.

The other general account of a virtue epistemology emphasizes more refined character traits. These are standardly called virtue responsibilist theories because they emphasize the responsibility that we think agents have toward their beliefs and, perhaps, a motivational component of virtue that may be necessary for some of the virtues. Some examples of the more refined character traits are conscientiousness, intellectual humility, or creativity, though this is not an exhaustive list. One consequence of the emphasis on agent responsibility or motivation for these kinds of character trait intellectual virtues is that some intellectual virtues are not necessarily truth conducive, and in fact, may produce a large number of false beliefs.

Neither the responsibilist nor the reliabilist deny the importance and existence of what the other emphasizes (i.e. responsibilists do not deny the importance or existence of the reliability of faculties like vision and memory, and reliabilists do not deny the importance or existence of more refined cognitive character traits); each emphasizes different aspects of their theories, though some virtue theories deny that some things like skills or faculties are virtues. Perhaps one’s preference for one view over another depends upon one’s intuitions regarding the nature of knowledge.

My own intuitions are divided on this issue. In some cases, knowledge is the kind of thing that involves the element of responsibility (and perhaps an associated motivation) as with epistemic conscientiousness. One example of this kind of case is scientific knowledge and the virtues associated with belief formation on the part of the scientist as epistemic agent. The scientist endeavors to find true, relevant propositions, and avoid false propositions, regarding her subject. An epistemically conscientious scientist does her best to do this because this is the epistemically right thing to do. In this sense, the scientist has a responsibility to believe in a certain way (e.g. accept certain propositions as true and reject others in light of the evidence). Furthermore, perhaps she is motivated to be epistemically responsible in this way because of her love of the truths of her subject. The knowledge that results from her inquiry, therefore, is virtuous.

Alternatively, perhaps there are some cases of knowledge where an account of virtue with reliability as the key element in knowledge is preferable. ‘Reliability’ means reliably, consistently getting to the truth. For example, in cases of sense perception reliability is the key feature for perceptual knowledge. One does not have to necessarily be motivated or aware of one’s motivations or responsibilities in forming the belief in order for it to count as knowledge. One just has to reliably form true perceptual beliefs (and avoid false beliefs). The advantage that reliabilist views have is that they emphasize the ‘success at getting to the truth’ component of simple cases of knowledge. This advantage is parallel to simple ethical cases where the primary component of evaluation is the state of affairs that results from the moral action (e.g. reduce the suffering of others).

Moreover, the two examples of virtues that I mentioned above (epistemic conscientiousness and reliable perceptual faculties) are examples of truth-conducive virtues. There are epistemic virtues that are not necessarily truth conducive. For example, the epistemic excellence involved with intellectual creativity is generally not thought to be truth-conducive. Consider the intellectual creativity involved with creating poetry. Though a poem may conform to strict structural rules, the elements of a poem (e.g. imagery or metaphor) which convey the meaning of the work do not convey literal truths, and in fact may lose their artistic meaning if taken literally.

In the next chapter, I am going to sketch and endorse a more narrow, reliabilist view of virtue epistemology, but I want to reiterate that my own intuitions regarding what counts as an intellectual virtue are that there is more than one kind of intellectual virtue. This is to say that I think that in some sense all of the above examples are robust examples of intellectual excellences, and therefore are intellectual virtues. Furthermore, for my purposes, I am dividing up the virtues into two kinds—truth conducive and non-truth conducive—though, there are other ways to divide up the virtues. Of the truth-conducive virtues, some of those are more like intellectual abilities or skills, while some of these virtues are more like character traits. Each of these kinds of virtues deserves recognition as virtues, and ideally, perhaps there will someday be a unifying account of all of these kinds of virtues, but that is beyond the scope of this dissertation.

Of the virtues that are specifically truth-conducive, the ones that are more like intellectual abilities or skills are often associated with simple cases of knowledge (as in perception) or reasoning. These virtues are best understood, I think, in terms of their reliable manifestations because when they are reliably manifested, they have the most value. My argument for this claim just is this. If getting to the truth is an epistemic good, then the reliable ability to do so counts towards what is good about arriving at and maintaining true beliefs. Getting to the truth is good, so the reliable ability to do so counts towards the goods associated with true beliefs.

For some of the truth-conducive epistemic virtues, reliability is not enough. These are the virtues that are less like abilities and more like intellectual character traits where reliability is not the most important feature of the virtue (though, it still may be a significant feature). My argument for the claim that for some of these intellectual virtues reliability is not enough is this. For example, reliable epistemic conscientiousness is a good epistemic character trait, but the reliability is not what makes being epistemically conscientiousness such a good thing (though, being reliable about being epistemically conscientious is a good). Epistemic conscientiousness is a good thing independent of whether or not one reliably gains a large number of true beliefs as a result of being conscientious (or, whether or not one is reliable about it). It is a good thing epistemically because it is the responsible way to form beliefs, and responsible belief formation is valuable. It is valuable because true beliefs are responsible beliefs.

Moreover, some of the character trait virtues are not truth-conducive. Creativity, if it is considered to be a character trait kind of intellectual virtue (rather than an ability), is one that does not necessarily aim at producing true beliefs. The goal of creative thought or action is to arrive at something new, and presumably the new creation is something of value. Because humans are cognitive beings, this virtue is rightly considered to be part of our intellectual makeup. The value in the creation is not merely the value associated with attaining the truth. Creative thought is valuable because of the new ideas that are produced and perhaps because of the ideas that are inspired in others as a result of the creation.[135]

These virtues have at least two things in common. They all have intellectual value of some kind. For example, reliable perceptual faculties allow one to form more true beliefs than false beliefs. Conscientious belief has intellectual value because the beliefs that are formed as a result of this particular character trait arise from the desire to seek the truth. Creativity has intellectual value because though the beliefs formed as a result of exercising this virtue are novel. Furthermore, all of these virtues are the kinds of things that are attributable to agents. Reliability, for anything, is attributable to that thing. For example, we attribute reliability to a thermostat that accurately displays the temperature on a reliable basis. The same is true for reliable perceptual faculties. We attribute reliability to an agent who reliably forms true perceptual beliefs, and this reliability gives more support to the belief that is formed. Conscientiousness involves desire—the desire to seek out true beliefs and avoid false ones. Desires are the kinds of things that are attributable to agents. Creativity is something that to my knowledge is only attributable agents.

[135] This is true for creative thought in the sciences as well. Suppose that intellectual creativity in the sciences is the intellectual excellence involved in coming up with new ideas, explanations, solutions, or hypotheses. This state of intellectual excellence may lead to true beliefs, but it may not. An intellectually creative scientist may come up with a novel hypothesis that turns out to be false, but the novelty of the hypothesis, itself, is still admirable, and the scientist is still intellectually excellent for deriving the hypothesis.

The reason why the two features above are important is that these features are central to credit theories of knowledge. While all theories of knowledge acknowledge that knowledge is something that is attributable to an agent, in general terms, a credit theory of knowledge emphasizes this aspect more strongly. There are two credit theories that I know of. First there is Greco's (2004) version that holds that reliable cognitive character is an important necessary condition for knowledge.[136] Second, there is Riggs' (2009) version that holds that the important necessary condition for knowledge is that the true belief is attributable to an agent as a cognitive agent.[137] I will go into more detail regarding these two theories of credit in the next chapter as I develop a particular version of VE, but for now I want to point out that the main difference between these two views is a matter of emphasis. Greco emphasizes reliability of cognitive character (probably because of his overall reliabilist stance), while Riggs emphasizes attribution to agents (due to his anti-luck considerations. Both agree that whatever knowledge is, it is a kind of intellectual achievement for an agent.[138]

Knowledge as achievement of something that is intellectually credit worthy is exactly what VE theories of knowledge have contributed to the overall discussion in epistemology. The notion of intellectual or cognitive credit sidesteps the abovementioned issues in epistemology, while at the same time is sometimes compatible with a particular view. Credit theories of knowledge do

[136] Greco (2003), op. cit., p. 12.

[137] Wayne Riggs, "Two Problems of Easy Credit" in *Synthese*, Vol. 169, No. 1 (2009), p. 203.

[138] Cf. Riggs, "Why Epistemologists Are so Down on Their Luck" in *Synthese*, Vol. 158, No. 3 (2007), pp. 329-344 and Greco's *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity* (2010), op. cit.

not necessarily resolve these debates, but shift the discussion to something more central.

First, regarding the foundationalism/coherentism debate, VE to my knowledge has not been able to resolve the debate between these two camps, but in shifting the direction of analysis from properties of beliefs to properties of agents as worthy/unworthy of intellectual credit for holding a belief, it changes the discussion in a fruitful way. The foundationalism/coherentism debate is a debate over the structure of justification. It treats beliefs as if they stand independent from a believer. The fruitful change that VE offers is to recognize that beliefs are intimately connected to a believer. Because beliefs do not stand independent from believers, properties of the believer and epistemic and external conditions under which beliefs are formed are important in contributing to whether or not a belief is attributed as justified. Whether the belief is produced through ability or motivation or responsibly the way in which it is produced determines the degree to which the belief is justified. Because this shift does not resolve the debate, VE is consistent with both foundationalism and coherentism.

Second, regarding the internalism/externalism debate, VE has not resolved this debate either, but this is an instance in which VE has advanced the discussion. The internalism/externalism debate is one of conflicting intuitions over the nature of justification. The debate is unresolved to be sure, however, the advance in discussion from VE here is to reinforce and reaffirm at least some of those intuitions. VE accounts of knowledge offer something for both internalists and externalists. VE accounts acknowledge that reliability, to some extent, is a necessary condition for knowledge, be it reliability at getting to the truth as an epistemic norm[139] or reliability as the successful manifestation of the virtue.[140] VE accounts offer internalists their necessary condition as well. Some kinds of knowledge (specifically, the non-simple kind as with perceptual knowledge) require internal constraints (in the form of the epistemic virtues) for justification.

Third, regarding the definition of knowledge, VE has contributed the most. The traditional definition of knowledge—justified true belief—is inadequate, as Gettier cases show. Since beliefs do not stand independent from a believer, it is a reasonable step to consider properties of a believer as part of the definition of knowledge. Various conceptions of VE have attempted different definitions, and indeed, at least one eschews necessary and sufficient conditions for knowledge.[141] The important advance in discussion from these accounts is that they make the intellectual character of the believer central to their accounts.

Conclusion

In this chapter, I have done the following. I have reviewed three of the major problems for epistemology in the 20th century. My aim was to show that VE is an alternative account of knowledge and justification that shifts the focus of analysis of these concepts. This shift is important because it changes the crux of the debate over issues like the structure and nature of justification and the definition of knowledge. In the following chapter I will endorse a specific view

[139] Cf. John Greco, *Achieving Knowledge: A Virtue Theoretic Account of Epistemic Normativity.* Cambridge University Press: New York, NY (2010), p. 3.

[140] Cf. Linda Zagzebski, *Virtues of the Mind*, Cambridge University Press: New York, NY (1996).

[141] Cf. Greco (2010), op. cit. p. 4.

of virtue epistemology. In the chapter after that I will conclude this dissertation with what I take to be the connection between VE and BCT.

CHAPTER 5

RELIABILITY, CREDIT, VIRTUE

Introduction

The central importance of the virtue approach to epistemology is its emphasis on the agent. Whether the emphasis is on the cognitive character of the agent or the cognitive ability of the agent, the approach is unique compared to previous theories. In those previous theories the agent as a knower is treated more like a placeholder for conditions on some abstract feature of the world, rather than the participant in the act of knowing. Of course, the agent plays a role in other theories of knowledge as well (e.g. S knows that p, always), however, VE's emphasis on the role of the agent as knower in a state of knowing (or, believer in a state of believing) showcases the role of the agent in the definitions of knowledge or justified belief. The aim of this chapter is to sketch a particular view of VE.

This chapter proceeds as follows. In Section 1 I briefly discuss my own dual conception of what an intellectual virtue is. For the rest of the chapter I will focus on one of these conceptions, intellectual virtue as ability or skill[142], in order to sketch the view more fully. In Section 2, I will discuss the role of reliability for this particular view and argue that reliability is a normative notion. In Section 3 I discuss the general notions of credit by exploring the familiar notions of athletic and moral credit. I do this in order to draw an analogy with the notion of

[142] I am using the terms 'skills' and 'abilities' interchangeably, though these terms sometimes have differing nuances.

intellectual credit. In Section 4 I conclude with a discussion of a particular virtue theoretic account in which the notion of credit is central to the account.

Section 1—The Nature of Intellectual Virtue

In this section I want to briefly discuss what I understand an intellectual virtue to be. I assume that a virtue, in general, is an excellence or a state of a thing that makes that thing good. An intellectual or cognitive virtue is an excellence of the intellect or cognitive state that makes the intellect good. In the last chapter I divided up the virtues into two categories—virtues that are more like skills or abilities to the extent that the manifestations of them tend to produce true beliefs, and virtues that are more like character traits that may or may not be truth-conducive. Undoubtedly, there are other ways to divide up and classify the intellectual virtues, and indeed, some[143] would not count certain abilities or skills (such as reasoning skills) as virtues. However, I think that dividing up the intellectual virtues this way is intuitive and for the most part uncontroversial.

To a first approximation, an intellectual virtue is similar to a moral virtue as defined by Aristotle. Roughly put, Aristotle defines a moral virtue as a state of character of an individual, under her control, that lies at the mean between two extremes. Some intellectual virtues are like this conception of moral virtues in that they are 1) states of an agent's intellectual character and 2) under an agent's control at least to some extent. When I say that intellectual virtues are states of intellectual character, I mean generally that they are excellences of an agent's intellectual makeup. By 'intellectual makeup' I mean the stable cognitive habits and dispositions that are part of an agent's intellectual life. I mentioned two

[143] Cf. Zagzebski (1996), op. cit.

examples of this kind of intellectual virtue in Chapter 4—intellectual conscientiousness and creativity. As I noted in Chapter 4, some of these character trait virtues are not necessarily conducive to true believing, as with creativity, while others are conducive to true believing, as with intellectual conscientiousness.

Furthermore, on this understanding, intellectual virtue is to some degree under an agent's control. This is to say that the virtue tends to be an acquired trait and may require agent motivation in order for the agent to develop this trait in specific cases. I say, "tends to be" an acquired trait because though virtues like creativity or conscientiousness may come naturally to the exceptionally gifted, most of us have to develop these kinds of virtues to some extent over time. Motivation is not a requirement for all cases of trait acquisition, though it seems like it is for some. For example, one could develop an intellectual character trait like creativity without a specific motivation in mind, especially if one is naturally gifted in the creative sense. In this case, creativity is a natural aspect of intellectual character, and though it may still need to be developed, because it is a natural aspect of intellectual character, motivation may not be required for development. However, it seems like for most people, motivation is required in order to develop a character trait (such as creativity). The development of the trait usually requires some work and practice, and this work and practice requires motivation. For example, an agent may have to be motivated to develop his intellectual creativity in order to complete a manuscript. The work and practice involved in the development of intellectual creativity also requires motivation.

Intellectual virtues are also much like other abilities or skills. Abilities in general are stable habits to achieve some end, and an intellectual ability is no different. While there may be more than one end to be achieved through intellectual ability, the most obvious one is to acquire true beliefs. This understanding fits easily with the virtues associated with perception. The abilities associated with perceptual cognition (e.g. sight, sound, touch, taste, smell) are intellectual virtues that require minimal training (if any at all) and do not always seem to be under the direct control of an agent. For example, when I look ahead and see a tree, I instantly recognize that it is a tree and form the belief that I see a tree. The belief that I form seems to require minimal effort and to be formed involuntarily.

These two ways of understanding intellectual virtue—as a state of character under the control of an agent and as an ability—are compatible and often overlap to some extent. For example, the character trait type virtues like creativity or conscientiousness are also abilities to some extent. The 'end' to be achieved is the creative result or the acquisition of more true beliefs than false beliefs, respectively. And some of the more 'ability' associated virtues can also be like character states in that their development and applications are under the control of an agent and may require motivation in order to develop. For example, while the ability to make good arguments, the ability to recognize reliable authorities, and the ability to evaluate evidence may come naturally to some, most of us have to develop these abilities over time. These abilities are under the direct control of the agent and may require motivation in order to develop. Furthermore, some intellectual virtues may only states of character (and not abilities at all), while others may be purely ability-like (i.e. and not character states of any sort). For example, open-mindedness may just be a character state, rather than an ability; and the virtues associated with perception do not seem like character states at all. This poses no special problem for the virtue approach. For, there is no need for there to be a single conception of the nature of an intellectual virtue. The important thing that both these conceptions of intellectual virtue have in common is that they are *excellences* of the intellect. All that's really essential is that we are able to recognize an intellectual virtue as such, and then can provide a fitting account of what makes it a *virtue* of the intellect.

From here on out in this chapter I will focus on intellectual virtues that fit the conception of an ability. A version of this conception of intellectual virtue lends itself especially well to Bayesian epistemology, and the purpose of this dissertation is to secure an important link between Bayesian epistemology and virtue epistemology. Moreover, because I take the emphasis on the *reliability* of the agent to be of particular importance in VE, I am going to focus on a version of reliabilism that emphasizes abilities in its account. If knowledge is of the kind Greco (2010) describes as a "success from ability,"[144] and I think it is, then reliability is central to developing and possessing an intellectual ability. The more reliable an agent is in exercising his ability, the more we attribute the ability to the agent. While I am sympathetic to the responsibilists, as described in Chapter 4, when specific kinds of knowledge or specific virtues are differentiated,

[144] Greco (2010), op. cit., p. 3.

reliability is uncontroversially important. One must be reliable in order for the character trait or ability to be of the kind that counts as *excellent*. For example, reliability is a necessary part of the virtue of open-mindedness, and of a virtuous perceptual ability as well.

Section 2—Reliability

In this section I will discuss a version of reliabilism that I think accurately captures the nature of knowledge. Previously, in Chapter 4, I discussed Goldman's process reliabilism, which is for the most part unsuccessful as a theory of knowledge. It is unsuccessful not only because of the generality problem, but also, and relatedly, because it is incapable of handling general, practical cases of knowledge. I said in Chapter 4 that the generality problem is a problem for all versions of reliabilism, so I will address the problem and a possible response later. For now I want to introduce a version of reliabilism that is more promising. This version is Greco's account of knowledge as success from ability (KSA). Greco's KSA account is as follows:

> S knows p if and only if S believes the truth (with respect to p) because S's belief that p is produced by intellectual ability.[145]

The use of the term 'because' here is intended to indicate a causal relationship. The idea is that S's belief regarding the truth with respect to p is caused in some way by S's intellectual abilities, skills, and powers.[146] Intellectual abilities and powers are understood here as the possession of intellectual virtues.[147]

[145] Ibid., p.71.

[146] Ibid.

[147] Ibid., p. 3.

Intellectual credit is deserved, and may be attributed to an agent when the agent believes p through the application of their intellectual ability.[148]

The KSA account is a reliabilist position. Reliabilism, broadly construed, is the view that emphasizes the truth-conduciveness of belief-formation via a process, method, or other epistemological factor. It is a view that applies to both theories of knowledge and justification. It is an externalist view regarding the issues of justification and knowledge because, for whatever species of externalism, the conditions for justification or knowledge are not limited to the mental contents of the believer. Because KSA is an externalist position, it faces the same objections faced by other externalist positions, including the generality problem.[149]

The emphasis on reliability supplemented with properly applied intellectual ability makes this view attractive as a theory of knowledge. 'Reliability', as used generally in epistemological contexts, refers to the idea that a process or an agent regularly gets to the truth, and this is a useful way to understand the notion. However, I want to draw attention to another, more common usage that is epistemologists tend to ignore. Reliability is a feature of a thing that makes it stable, and when applied to a process or procedure, it is a feature that makes the process or procedure repeatable. Stability and repeatability are valuable in general. The stable disposition of an agent, for example, is one on the basis of which we can predict future behavior. A repeatable method in

[148] Ibid., p. 141.

[149] In Chapter 4, I discussed some of the general objections to externalism.

science, for example, is valuable because results from the method can be rechecked, and thus verified or refuted.

Reliabilism, by itself, is an attractive feature of a theory of knowledge or justification because it emphasizes an important truth-conducive property of belief formation—reliability. Reliability is important in the formation of true beliefs because it adds the value from stability and repeatability to the belief that is formed. It is uncontroversial that true beliefs, justified true beliefs, and knowledge are all valuable, and we generally think of these as valuable in ascending order. One source of this value is the reliable means by which the belief is formed.

My claim that one source of value is reliability, though, is not uncontested. Zagzebski (2003) argues that reliability *per se* has no value or disvalue. She uses several examples to make her point. The goodness of the espresso does not come from the source that produces it, the espresso maker. The espresso would be just as good whether it was produced by a reliable or unreliable espresso maker. A reliable dripping faucet is not good, not because it reliably drips, but because dripping water is not good.[150] These examples are supposed to be analogous to true belief and the source of true belief (as in a belief forming process in an agent). I agree with her that reliability alone does not account for the goodness that true beliefs have. True beliefs, whatever value they have, are valuable regardless of whether or not they are produced from a reliable process. For example, the belief that I form when I look ahead and correctly identify the tree

[150] Linda Zagzebski, "The Search for the Source of the Epistemic Good" in ***Metaphilosophy***, Vol. 34, Nos. 1-2 (January 2003), p. 12.

that I see before me is valuable (has value) regardless of whether or not my belief forming process is reliable.

However, I disagree that reliability can have no value *per se.* If reliability is understood as stability and repeatability, then as applied to some kinds of cases, reliability certainly does add value. For example, while it is true that a reliable espresso maker does not add value to the espresso (the espresso is just as good or bad whether or not it is produced from a reliable maker), reliability in the ability to produce good espresso makes the espresso maker valuable. In other words, the espresso maker itself has value because of its reliability to produce good espresso. Though dripping water from a faucet is not good, the fact that the faucet drips reliably makes the faucet not good. Reliability itself confers value on the thing that is reliable. So, a reliable belief forming process is valuable because of its reliability in getting to the truth (at least).[151] Reliability may not explain the value *simpliciter* of a true belief, but if the belief is formed from some reliable faculty, method, or process, then the belief has some added value in virtue of the fact that it was formed by a reliable (thus, valuable) process. The same holds for true beliefs that are unreliably formed. Of course they have some positive value because they are true, but they also seem to have some negative value as well. For example, suppose I form the belief that *the winning Powerball numbers are 4, 21, 20, 12, 26, 22* as a result of seeing these numbers on the back of a fortune from a fortune cookie. Suppose further that this belief turns out to be true. Now, this true belief was formed from an unreliable process, but is valuable because it

[151] As discussed in Chapter 4, some reliable belief forming processes may not produce true beliefs, but may still be considered valuable. For example, a belief forming process that produces new ideas (though most of the beliefs that the process produces are false) is still valuable.

is true; thus it does not have the substantive value that reliably formed beliefs have. For example, a belief formed from my ability to reason well, whether it is true or not, has more value than the fortune-cookie belief.

Reliably formed justified true beliefs are good things to have, and I have argued that they are good at least in part because of the truth-producing reliability of the process by which they are formed. But reliability is not enough for a robust theory of justification or knowledge. Reliability is not enough for a robust theory of justification because although a reliable method, process, or ability may confer some amount of justification (e.g. the belief is justified in virtue of the fact that it was produced by a reliable process), and therefore value, it does not confer the full degree of justification that seems to be required for knowledge. Take for example a justified true belief that p formed by my reliable faculty of sight. The reliability of my faculty alone cannot be the sole source of my justification for believing that p, though reliability does confer some justification.

The reason why reliability alone does not confer the kind of justification required for knowledge is that there are many instances where one can reliably form a true belief with some justification, but nevertheless cannot legitimately claim to know. For example, suppose I go to only a partially reliable source like Wikipedia to find out whether or not Brasilia is the capitol of Brazil. Wikipedia is the kind of source that is only partially reliable. The belief that I form, *that Brasilia is the capitol of Brazil*, is justified, but not to the degree that is required for knowledge. To have sufficient justification for the belief to count as knowledge, I would need something more. Thus, reliability alone is also cannot sufficient for a theory of knowledge. Gettier cases show that a belief can be produced by a reliable belief forming process (e.g. inference), yet fall short of knowledge because the justification conferred falls short of what's needed for knowledge.

Reliability is a feature of justification and knowledge that is the same for both those who come by it naturally and those who can only acquire it via hard work. So although acquiring something through hard work may deserve some special kind of credit (and virtue theories tend to draw on credit-worthy attributes), reliability is itself credit-worthy regardless of how it is acquired. To see the point, consider that some people are naturally better at solving mathematical problems than others. They seem to get truths of mathematics without nearly as much study or work as others require. However, the fact that the ability to solve mathematical problems comes naturally for them does not detract from the value of their ability to get correct results. In other words, those who are naturally equipped with mathematical ability deserve no less credit for their mathematical beliefs than those who have to work incredibly hard to acquire similar abilities. The feature of the ability common to both the agent who comes by it naturally and the agent who has to acquire it via hard work is reliable.

Section 3—Credit

The notion of the *credit-worthiness* of attributes and abilities must play an important role in a virtue epistemology. Credit is the kind of thing that can be deserved, and most importantly, the kind of thing that can be attributed. When we attribute credit to an agent for an action or an outcome, we do so because we think an agent deserves it. For example, when a professional basketball player makes a shot from the three point line, we give him credit for making the shot. When a student answers correctly on an exam, we give her credit for giving the correct answer. The means by which someone achieves (or deserves) credit is through their ability, whether it is a motivated effort or not. This suggests that the notion of credit may play a significant role in a virtue epistemology, since credit attends virtuous ability.

When I say that the means by which someone achieves (or deserves) credit is through their ability, I mean the following. Since abilities are the stable habits to achieve some end, they are the kinds of things that are more resistant to accidental features of a situation that may either accidently bring success or prohibit achievement. From here on out I'll call such accidental features "luck". For example, when one has the ability to make a shot from the three point line on a basketball court, this means that one has the stable habit, entrenched in some way, to make that shot, whether or not one is motivated to make the shot, and whether or not one has ever made that kind of shot before. So, in some sense, even a non-expert, non-professional has the ability, though this sense is very weak. In this case, the case in which the non-basketball playing novice makes the shot, the novice has achieved the end, and therefore, is worthy of some credit, but not a lot of credit. The novice certainly deserves less credit than that professional. The reason why the novice deserves less credit is because she lacks in ability, and her success is due more to the lucky features of the situation than to stable habits.

It may seem controversial to say that a novice deserves any credit for an act that produces an outcome that is more due to the accidental features of a situation than her ability. The reason why the novice deserves credit (though it is minimal) is because the result that occurred came about as a result of her action. This kind of credit, though, is negligible. Consider Greco's example of athletic credit, Bucky Dent's homerun that eliminated the Red Sox from the play-offs in 1978.[152] Bucky Dent was not a homerun hitter; indeed, he hit very few homeruns over the course of his career. However, on this particular occasion, he managed to hit a homerun. Given his past hitting record and the conditions on the field at the time, it was a remarkable occurrence.[153] Dent received credit for hitting that homerun, as did the Yankees as a team. But, as Greco points out, Red Sox fans do not give very much credit to either Dent for the three run homer or to the Yankees for their subsequent win. Red Sox fans attribute the homerun to mere *bad luck* because accidental features of the situation, such as conditions on the field, had more influence on the outcome than did Dent's hitting ability.

In moral action, the same notion of credit can be applied. Riggs puts it this way:

> Someone who produces some morally good end by pure luck deserves less moral credit than does someone who produces the end by application of her skills and abilities. These kinds of considerations **are** what lie behind our intuitions about cases of "moral luck." The drunk driver who fails to

[152] Greco(2003), op. cit., p. 122.

[153] So much so, that even Dent could not believe that he had made the homerun. Dent is quoted as saying, "I knew I had hit it high enough to hit the wall…but there were shadows, and I didn't see the ball land. I didn't know I had hit a homer until I saw the umpire signaling. I couldn't believe it." in The New York Daily News, June 10, 2003 available at http://articles.nydailynews.com/2003-06-10/news/18226674_1_new-york-american-league-wrong-place-bucky-dent/2

> run over someone only because she encountered no pedestrians on the way home deserves little or no credit for that positive outcome (or lack of a negative outcome) because this outcome was not (sufficiently) brought about by the application of her abilities, powers, and skills, but rather due to chance alone (lack of pedestrians).[154]

The drunk driver is "morally lucky" to the extent that because of external circumstances in her situation, she does not injure pedestrians during her drive (the good consequence). The common intuition is that she does not deserve much credit for this. It is merely an accidental feature of the situation that there were no pedestrians on the road at the time of her driving.

In the moral realm, actions are more morally praiseworthy (i.e. morally credit-worthy in Riggs' sense) when they can be attributed to the powers, abilities, actions, and decisions of the agent who brings about the consequence. A drunk driver and a sober driver may achieve the same end of not injuring pedestrians. In some sense, both can be attributed some positive credit, while the sober driver, obviously, deserves more positive credit. The drunk driver is less capable in her driving abilities than the sober driver, so it is merely due to luck that the good end (not injuring pedestrians) comes about for the drunk driver.

Here is an example of negative credit-worthiness, adapted from a famous example due to James Rachels.[155] Both Smith and Jones stand to inherit a large amount of money upon the death of their young cousin. In Smith's case, he sneaks into his cousin's bathroom while his cousin is taking a bath, drowns him,

[154] Wayne Riggs, "Reliability and the Value of Knowledge" in *Philosophy and Phenomenological Research* Vol. 64, No. 1 (Jan., 2002), pp. 92-93.

[155] James Rachels, "Active and Passive Euthanasia" in *Contemporary Moral Problems*, James E. White (ed.), Wadsworth: Boston, MA (2012), p. 156.

and makes the surroundings appear as if the child died accidently. In Jones' case, he, with the same intention as Smith, sneaks into his cousin's bathroom while his cousin is taking a bath. Just as Jones enters the bathroom, his cousin slips and begins to drown. Jones does nothing to save his cousin while he watches his cousin die. We consider both Smith and Jones to be both morally blameworthy. Smith takes the direct action to kill, while Jones takes no action. Rachels' point is that while there is a bare difference in the facts of each situation, there is no moral difference. The prevailing intuition about Jones is that he is as morally blameworthy for the death of his nephew as Smith is.

I contend that the reason why there is no moral difference between Smith's case and Jones' case is because while luck plays a role (i.e. the nephew slips), from the moral perspective, the outcome is in fact due to the agent's powers, abilities, actions, and decisions. Though Smith takes direct action to kill his cousin and Jones takes no action, we find both Smith and Jones equally morally blameworthy because both intended to act in such a way as to bring about the death of their respective cousins. The "lucky" features of Jones' situation, though, allowed for Jones to kill his cousin with inaction. In fact, his decision to do nothing to keep his cousin from drowning together with his resulting inaction, are morally significant facts of his situation. His inaction is what makes him "credit worthy" in a negative sense—i.e. blameworthy.

For both kinds of credit discussed so far, athletic and moral, notice that credit need not be an absolute, all or nothing attribution. Credit is deserved and attributed in degrees. When an expert homerun hitter hits a homerun, more credit is attributed to her for the homerun than attributed to a novice who hits a homerun. While an expert homerun hitter deserves credit for hitting a homerun, the circumstances under which the homerun is hit do still play a role, and even then credit is attributed in degrees. For example, when an expert hitter hits under the “right” conditions (i.e. no wind or favorable winds, lower air pressure, lower humidity), the circumstances for achieving the homerun are favorable for success because the ball travels through atmospheric conditions with less resistance. Perhaps less credit is deserved for the homerun in this instance, even for the expert. When an expert hitter hits under less than optimal conditions (e.g. strong headwind, high air pressure, high humidity), atmospheric conditions offer more resistance against the ball, making the home run less likely. Perhaps more credit is deserved for success in this instance.

The same is true for moral credit. Suppose Smith is a sober driver who must drive through an area dense with pedestrians and Jones is a drunk driver who must do the same. When Smith is successful at navigating pedestrian traffic, she deserves more positive moral credit (moral praise) than would Jones who also successfully navigates through pedestrian traffic. Furthermore, Smith also deserves more credit in this case than she would in the case where there the pedestrian traffic is sparse.

The examples of athletic credit and moral credit suggest analogous credit-worthiness assessments for cases of epistemic or intellectual performance. An agent deserves more or less epistemic credit, depending on the degree to which her powers, abilities, skills, and (where appropriate) motivations are involved in bringing about her true belief. For example, a student who makes a correct guess at the answer deserves less credit for her belief than a student who arrives at the answer through her abilities, skills, etc. When an agent believes the truth with respect to p as a result of her abilities, skills, etc, she has, to borrow Greco's term, *achieved* true belief.

Furthermore, if I am right about reliability being crucial to accounts of justification and knowledge, then the notion of credit for believing truly is crucial as well. There is an interplay between the reliability of an agent and the extent to which credit is due to the agent for holding the belief. For, if an agent happens to believe truly, but via an unreliable process, then little credit is due. Thus, on a virtue account in which knowledge is taken to be "true belief achieved in a credit-worthy way", the belief may fail to count as knowledge. However, on this kind of account of knowledge, when an agent believes truly through the application of reliable truth seeking processes, then the kind of credit due to the agent may raise the belief to the status of knowledge. Thus, on this kind of view, 'knowledge' is an honorific for an agent's beliefs when they are true and acquired via the right kind of process involving the right kind of intellectual abilities (which may be either natural or acquired abilities).

Section 4—Virtue

What is the connection between epistemic credit as described above and intellectual virtue? In this section I will attempt to answer this question. My preferred account of knowledge and justification, Greco's account of knowledge as success from ability (KSA), incorporates the idea that whatever knowledge and justification are, they are attributable to an agent. So the KSA view is a virtue-theoretic account that takes credit for a belief into its account. This is in contrast to the standard definition of knowledge as *justified true belief*, which permits a rather weak connection to agency. On the usual account a belief need merely be creditable to the agent as having been justified for him (or by him). This weak connection to agency in the traditional account of knowledge as *justified true belief* is one reason that the traditional accounts are so *easily susceptible* to Gettier problems, in part because the justificatory process may fail to be one that reliably produces true beliefs for that agent under the present circumstances.

Briefly, Greco's KSA account is as follows:

> S knows p if and only if
>
> > S believes the truth (with respect to p) because S's belief that p is (reliability) produced by (virtuous) intellectual ability.[156]

The use of the term 'because' here is intended to indicate a causal relationship. The idea is that S's belief regarding the truth with respect to p is caused in some way by S's intellectual abilities, skills, and powers.[157] Intellectual abilities and powers are understood here as the possession of intellectual virtues.[158] Intellectual credit is deserved, and may be attributed to an agent when the agent believes p through the application of their intellectual ability.[159]

[156] Greco (2010), op. cit., p.71.

[157] Ibid.

[158] Ibid., p. 3.

[159] Ibid., p. 141.

The KSA account is a reliabilist position. Regarding the issues of justification and knowledge it is an externalist position.[160] I take the emphasis on the reliability of the knowledge acquisition process, together with the point that reliability is insufficient for knowledge, but must be supplemented with properly applied intellectual ability, to be an especially important feature of the KSA account. The point is that although some agent *may* reliably form true beliefs, that agent would fail to be creditable for having *knowledge* if that process failed to draw on his intellectual ability (e.g. if he were a very reliable guesser).

Reliability is central to an agent-centered account of justification and knowledge, so it is a requirement of any adequate epistemology. This is because an important feature of knowledge should be a kind of stability that attends reliability. In other words, although our justifications of particular propositions may change (thus changing their status as *knowledge*), conceptually for a proposition to count as *knowledge* it should be the kind of thing (have the kind of epistemic properties) that permit it to be reliably available *from* a *knowledgeable agent*. If an agent *reliably believes the truth*, then that agent should be is a stable source of knowledge.

Furthermore, an agent's intellectual abilities and powers should properly be part of that person's cognitive character or make-up. Thus, it should be dispositional. In other words, an agent has a cognitive ability or power as part of her (at least somewhat) stable cognitive makeup, which disposes her to believe a

[160] This is compatible with the historical reasons I presented in Chapter 4. The issues that arise in the historical debates presented in Chapter4 are not entirely resolved by a virtue-theoretic approach. However, the direction of analysis changes from properties of propositions to properties of agents, and this appears to be an important step in the right direction for a resolution of traditional issues.

proposition only when that belief is formed via certain reliable processes that draw on the right sorts of intellectual abilities. Epistemic ability is the kind of thing that is stable, and that lends stability to the knowledge possessed by an agent. Furthermore, when an agent has an ability that is a stable disposition towards achieving some particular kind of end, then that agent may justly be credited for achievements that result from that ability.

The value of the stable disposition is the same, at least in some sense, whether the ability is natural or acquired. Of course, we may value an ability that we have acquired through hard work (rather than via nature) because we had to work hard for it, but this is a different sense of the term ‘value’. The sense of ‘value’ to which I refer, though, is the value that the ability has due to the stable disposition or dispositions towards achieving valuable ends that make up the ability. For example, the value of the ability to reliably form true mathematical beliefs is the same whether or not it is easily acquired. Moreover, whether or not the ability is easily acquired is irrelevant to attributing credit.

Again, the notions of reliability and credit are import to VE. What makes someone worthy of credit is their reliability. The more reliable an ability is, the more credit to be attributed for the successful achievement through ability. To get the idea, consider the Dent example again. Dent was not a reliable homerun hitter. His homerun was due more to the ‘lucky’ conditions for him on the field. Though, Dent gets some credit for hitting the homerun (contra Red Sox fans), because he is an unreliable hitter, he does not get much credit. A reliable hitter, deserves more credit, even if conditions for him are ‘lucky’ at the time of success.

This kind of account of knowledge as credit recognizes, as Greco notes, the kind of normativity that comes via reliable success from ability, and as such is a kind of externalist account of knowledge.[161] In any area of action (e.g. moral, athletic, intellectual), reliable success from ability is positively evaluated. Similarly for reliable failure or inconsistency of action—we both are evaluated negatively. In Chapter 2 I argued that reliability is a normative constraint on acquiring rational belief. I agree with Greco that when an agent is reliable in believing truly, that reliability satisfies a central kind of epistemic standard. Reliability may not be the only epistemic standard, but on the present account of credit it is the most salient standard.[162]

Greco proposes the following theory of intellectual credit attribution. S deserves intellectual credit for believing the truth regarding *p* only if:

a. believing the truth regarding *p* has intellectual value,
b. believing the truth regarding *p* can be ascribed to S, and
c. believing the truth regarding *p* reveals S's reliable cognitive character.

Alternatively:

> S's reliable cognitive character is an important necessary part of the total set of causal factors that give rise to S's believing the truth regarding *p*.[163]

I think that this theory of intellectual credit attribution gets things right regarding the notion of epistemic credit. Believing the truth regarding a

[161] Ibid., p. 7.

[162] Another normative constraint on rational belief that I have in mind is consistency. There are different senses of consistency, most notably deductive and probabilistic. I take consistency to be independent of reliability.

[163] Greco (2010), op. cit.

proposition usually has at least some epistemic value. When the belief is creditable to an agent, the minimal condition on knowledge or justification is met. But, most importantly regarding epistemic credit (for concepts like knowledge or justification), an agent's cognitive character has to play a relevant role in the reason why an agent believes the truth. For instance, believing that it is not raining (when it is in fact not raining) is epistemically valuable. It is a good thing to believe the truth about the external world. When that belief is held by an agent, the belief is creditable to that agent. But more importantly, when an agent believes as a result of her cognitive abilities, the agent deserves credit for holding that belief because her positive intellectual abilities are part of the causal reasons for her belief.

Epistemic credit seems to me to be central to the concepts of knowledge or justification in general. Evidence, reasons, arguments, etc. offer justification for propositions. However, none of these are relevant without there being an agent who holds the corresponding beliefs about evidence, reasons, or arguments. Furthermore, the mere fact that an agent might justifiably hold the belief is not enough, as Gettier cases show. The belief has to be attributable to an agent due to her intellectual abilities entering into a causal process in the right way for her true belief to count as knowledge.

Finally, I take epistemic credit to be something that admits of degree, like other forms of credit. Dent may be due some credit for the homerun he hit. However, the amount of credit he was given depended on who was evaluating him. Red Sox fans gave him little to no credit—they took the accomplishment to be almost entirely attributable to luck rather than ability. Yankees fans, though astonished, give Dent more credit. What the Sox fans and the Yankees fans disagreed about was the extent to which Dent's abilities contributed to his success. For, Dent was in fact trying to get a hit; and most homeruns happen that way (usually a player who hits a homerun is not specifically *trying for a homerun*). And Dent did have fairly good hitting ability. So, arguably, he should get as much credit as most homerun hitters. In any case, I think that epistemic credit works in much the same way. When someone comes to justifiably hold a belief, then he should get some credit. The amount of credit depends on the degree to which the agent's abilities or powers properly contribute to the agent's holding the belief.

CHAPTER 6

EPISTEMIC VIRTUES, PRIOR PROBABILITES, and NORMS for HUMAN PERFORMANCE

Introduction

Very little work has been done to try to connect Bayesian epistemology with any version of traditional epistemology, let alone virtue epistemology.[164] The purpose of this dissertation is to try to do just that. In Chapter 2, I discussed Bayesian Confirmation Theory (BCT) as a way of providing the logic and epistemology of scientific inference. For the most part, BCT is usually conceived of as providing an idealization of warranted belief strengths for real agents. In Chapter 3 I discussed one recent application of the principles of BCT to cognitive science, *rational analysis,* which is less idealized in that it is an empirical project that shows the extent to which real human agents may reason in a way that conforms to Bayesian principles. In Chapter 4, I discussed the historical motivations for VE. In Chapter 5 I presented a brief sketch of a reliabilist version of VE. In this chapter I will propose and discuss some natural connections between the Bayesian epistemology and VE.

I think of VE as being less idealized than BCT, but still somewhat idealized, as normative models tend to be. VE is, in a sense, an empirical project in that it looks to the instances of epistemological success for normative standards. By *epistemological success* I mean an agent's success associated with

[164] I know of only one other person working on a similar area. Sarah Wright from the University of Georgia recently presented "Partitioning Virtuously" at the 2011 Central Division APA. Her paper focuses on applying a virtue approach to explain how we reason well about probabilities. I did not attend the Central APA, but have had correspondence with Dr. Wright regarding her work.

epistemic acts of thought, for example, justified belief, acquiring true beliefs, acquiring knowledge, thinking creatively, etc. In this sense VE is agent based, and both VE and BCT are agent based conceptions. By *agent based* I mean that the conceptual priority is given to particulars about agents and the conditions under which epistemic concepts like knowledge, justification, or confirmation apply with respect to individual agents.

Section 1—Why Bayesian Epistemology Needs Something Else

The main goal of BCT is to provide a logic and epistemology for how we arrive at strongly confirmed theories that are true. As a logic, and to some extent as an epistemology, it is an idealization. Additional normative constraints are surely relevant, not only to belief in general, but also to beliefs that arise in scientific contexts. Traditional epistemology mainly endorses these additional normative constraints via analyses of the concepts of justification and knowledge. Though the approach from traditional epistemology is compatible with Bayesian epistemology, the primary focus of Bayesian epistemology has been on other concepts, such as belief strengths and degrees of confirmation, which are ultimately represented by probabilities.

The majority of Bayesians are subjectivists who take Bayesian probability functions to represent the belief strengths of ideal agents, where belief strengths are measured on a probabilistic scale. The belief strengths of ideal agents are probabilistically consistent—that is, they satisfy the laws of probability theory. The belief strengths of real agents should satisfy the same laws of probability as well (i.e. must be probabilistically consistent), on pain of a kind of pragmatic flaw in how they rank possible actions to perform when outcomes are uncertain. An agent whose belief strengths violate the probabilistic laws is subject to select "best actions" that guarantee a loss, regardless of how the world may turn out—her decision are subject to a so-called Dutch-book. Furthermore, decisions to act based on probabilistically consistent belief-strengths can always avoid this kind of sure-loss. Thus, subjectivist Bayesians use probabilistic consistency as a standard of rationality. They often call probabilistically inconsistent agents *irrational*.

Any broader epistemology should accommodate this point about decision and action. This is to say that any robust theory of knowledge should include a comprehensive theory of decision that links belief and knowledge to decisions regarding best actions. The only comprehensive theory of this kind that we have at present is Bayesian decision theory. So any epistemology that pretends to be comprehensive should either incorporate Bayesian decision theory or be ready to replace it with another theory that fulfills the same role of linking belief and knowledge to decisions regarding best actions. To my knowledge there are no candidate replacement theories of decision available.

The subjectivist Bayesian account, though, does not supply an adequate epistemology on its own. Its account of rationality is merely that of probabilistic consistency. As we saw in Chapter 4 in the discussion of coherentism as a theory of the structure of justification, mere consistency, even if it is probabilistic consistency, is not enough for an adequate epistemology. Recall, for instance, the objections to coherentism that points out that a piece of fiction can be completely consistent, although each of the propositions in the story is false. In the same way, belief strengths may be probabilistically consistent (i.e. not violate probabilistic laws), but not at all closely aligned to the truth (e.g. it is probabilistically coherent for an agent to have degree-of-belief .95 that Elvis is alive on Mars, provided the agent also has, for instance, degree-of-belief .05 that Elvis is not alive on Mars). The subjectivist Bayesian's account of probabilistic consistency has nothing to offer in the way of getting to the truth via evidence.

Objectivist Bayesians differ from subjectivist Bayesians with regard to how they interpret the kinds of probabilities involved in Bayesian confirmation functions. Some take confirmational probabilities to be relative frequencies, some take them to be objective chances.[165] Other objectivists take Bayesian confirmational probabilities to be strictly logical in some sense. Historically many Bayesian logicists have taken the logically specified prior probabilities required by the Bayesian formalism to be determined by some form of a principle of indifference.[166] This is in contrast to the subjectivist Bayesians, who allow the prior probabilities to represent an agent's degree of belief in a hypothesis prior to taking the evidence into account.[167] Sometimes there is agreement among various kinds of Bayesians with regard to how evidential support (i.e. confirmation of hypotheses) is supposed to work, sometimes not. Their differences are especially

[165]J. Williamson. *Objective Bayesianism with Predicate Languages. Synthese* **163** (2008), pp. 341, 343.

[166] Earman, ibid, p. 197. A generalized version of the Principle of Indifference says that when there are n mutually exclusive and exhaustive outcomes, one should assign the probability of 1/n to each outcome. There are other, more recent variations on the Principle, but for simplicity's sake we can focus on this one for now.

[167] William Talbott, "Bayesian Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epistemology-bayesian/>.

strong when it comes to the assessment of prior probabilities for hypotheses.[168] However, arguably, any version of Bayesianism that pretends to be a basis for an epistemology of evidential support for scientific hypotheses and theories should depend only on objective or rational factors to figure into an agent's assessment of prior probabilities.

This kind of divide between subjective and objectivist Bayesians also applies to the logic of confirmation. The subjectivist takes confirmation functions to be probability functions that represent belief strengths of ideal agents. A conditional probability for a real agent represents a belief strength an ideal agent would have in a hypothesis on a body of evidence if that evidence represents all she knows (with certainty) that is relevant to a hypothesis. The objectivist (or logicist) takes confirmation functions to be conditional probability functions. Although traditionally many objectivists have used the principle of indifference, to assign prior probabilities, that is not really essential to the logicist line. What is essential is that logicists take the numerical values of the conditional probability functions to represent a measure of argument strength—the measure of how strongly evidence supports a hypothesis. Indeed, the logicist may completely part company with the subjectivist Bayesians. The logicist is only interested in a logic of confirmation, not in and account of decision. To the extent that the logicist wants to tie the confirmation function into the logic of decision, the logicist takes confirmational probability functions to be completely distinct from the degree-of-belief functions used in decision theory. Rather, the logicist holds that degrees of

[168] I discussed the differences between subjectivist and objectivist Bayesian interpretations of probabilities in Chapter 2.

belief should be informed by degrees of confirmation – that the confirmation function is distinct, but should inform what the agent believes, and how strongly she believes it. Thus, the logicist Bayesian approach to confirmation does not pretend to provide anything other than a logic of evidential support. So, if the logic of Bayesian confirmation theory is to do any real epistemological work, it must be embedded within a broader theory of knowledge. Furthermore, even subjectivist Bayesian confirmation theory fails to provide anything like an adequate epistemology, since arguably it does not, on its own, provide an account of evidential support, but only an account of probabilistic coherence (consistency) appropriate to decision theory.

Perhaps the most telling argument that subjectivist Bayesian belief functions cannot be Bayesian confirmation functions come via the so-called problem of old evidence. To see how this argument works, first notice that the primary role of a Bayesian belief function is to represent an idealized agent's belief strengths in various propositions (or sentences) given everything she knows. This is because the belief strength measure is supposed to combine with the agent's desires for outcomes in the world (represented by her utilities) to provide decisions on actions to be taken. A belief function that falls short of representing her full belief strengths for possible states of the world cannot properly play this decision theoretic role. However, a confirmation function is supposed to measure the support of evidence for a hypothesis based on how likely the hypothesis *says* that evidence is, regardless of whether the agent knows that the evidence claim has turned out to be true.

For example, when a coin has turned up “heads” and the agent knows that, her belief function should say that her belief strength in “heads” is 1. It is a theorem of probability theory that when Pr(E) = 1, Pr(E | H) = 1. So, this agent cannot use her belief function P to represent that the hypothesis, H, says that the coin is fair—i.e. her belief function, P, cannot have it that P(E|H) = ½. So, her belief function cannot represent what a hypothesis *says* about evidence. So, her belief function cannot properly measure the confirmation of hypotheses on the evidence. This is one aspect of the so-called *problem of old evidence*. It shows that a Bayesian belief function cannot be the same thing as a Bayesian confirmation function.

The fact that belief functions cannot be confirmation functions leaves a gap in the subjectivist Bayesian account of the logic of confirmation. Subjectivist Bayesians have an account of ideally consistent belief strengths, but no good account of a confirmation theory that gives rise to these belief strengths, much less an account of how evidence should inform belief in a way that makes true claims come to be strongly believed. Subjectivist Bayesians must appeal to some separate account of confirmation for that. If that separate account is also Bayesian, it must employ confirmation functions that are distinct from the Bayesian belief functions that are used in Bayesian decision theory to represent the belief strengths on which decisions are made. So, even Bayesians in the subjectivist camp must draw on a separate logic of confirmation functions that represent a kind of logic of evidential support. In addition, they need an account of how Bayesian support (Bayesian confirmation) is supposed to inform Bayesian belief strengths. One simple version of this account would be that an agent's belief strength for a hypothesis, H, should be updated to have the value of her confirmation strengths based on all the evidence she knows (with certainty). However, this remains far from a full theory of knowledge.

Section 2—More for Bayesians to Consider

On the account given in Chapter 2 belief is not binary, not all or nothing. It comes in degrees. This approach to belief has previously been a point of conceptual disagreement between traditional epistemologists and Bayesians. Traditional epistemologies focus on belief as a binary categorical concept. So, this area of disconnect should be a starting point for discussion. There is no doubt that real agents are more confident in some propositions than others, while at the same time they hold certain beliefs in the categorical sense. Idealizations, to the extent that they are useful and accurate depictions of real agents, should share this same feature. Rather than see this feature, degrees of belief, as a flaw for humans as epistemic agents (as the model of binary belief suggests) to be overcome or ignored in the idealization, this feature should be incorporated into the idealized model.

Although it is convenient to characterize belief strengths for the ideal agent in terms of probabilities that have precise numerical values, a real agent will usually have a difficult time determining exactly to what degree she believes in the existence of a god, or that the universe is 13.7 billion years old. However, the model of idealized agents and their levels of confidence does not really need to rely on numerical values. An alternative approach is to employ a kind of qualitative probabilism that takes comparative confidence relations to be basic (e.g. relations like *the agent is more confident that B than that C*). On this model binary belief may be represented as comparative confidence above some appropriate threshold.[169]

I will not fully address the issue of how degrees of confidence and qualitative comparative confidence are related to binary belief. A few remarks on the issue will serve my purposes.

The qualitative approach to comparative belief seems a much more natural model of belief for real agents than does the numerical degree of belief idea. Indeed, the notion of numerical degrees of belief can be shown to derive from the qualitative comparative notion. The numerical notion can be shown (via a representation theorem) to be an overly precise idealization of the qualitative comparative notion (the notion that the agent *is at least as confident that B as that C*). Intuitively plausible rules for this comparative notion have been developed and explicated.[170] These rules for comparative confidence (e.g. that comparative confidence is transitive) are normative constraints that real agents may in fact violate. (This kind of logical norm is specified in terms or axioms or rules, whereas most epistemic norms with which I am concerned in this project are less formal.) Thus, there are at least two important kinds of belief that are epistemically important: the categorical doxastic attitude that is binary, admitting of no degrees; and the comparative notion, which may be representable by

[169] Hawthorne, “The Lockean Thesis and the Logic of Belief” in Franz Huber and Christoph Schmidt-Petri (eds.), *Degrees of Belief*, Synthese Library, 342, 2009, Springer, pp. 49-74.

[170] Ibid. These are adaptations of the axioms qualitative probability.

numerical degrees. Insofar as the binary notion is supposed to satisfy the norm of logical consistency—that the agent's whole body of beliefs should be logically consistent—it is every bit as much of an idealization as the comparative notion (with its axioms or rules for coherent comparative confidence relations to follow).

My main concern is not these idealizations as such, but the epistemic life of real agents, and how these norms are supposed to benefit real agents. The point of introducing the idealized model is to provide easily recognizable norms of belief formation and assent for real agents to follow. The normative model provides a kind of exemplar that shows us what the normative standards are.

In Chapter 5 I endorsed a view of knowledge and credit in which intellectual character plays an important causal role in tending an agent's belief towards truth. On the view presented in Chapter 5, one of the necessary conditions for knowledge and justification is that it be credit worthy. Just as confidence in propositions comes in degrees, there is no doubt that creditworthiness comes in degrees as well. If intellectual credit is closely analogous to other kinds of credit (moral or athletic), then the degree to which someone deserves credit for true belief depends on the degree to which believing truly can be attributed to the agent's skills or abilities.

Section 3—Confidence, Belief, and Acceptance

Mark Kaplan (2002) argues for a modest version of subjective Bayesianism in which he adds to the notion of degrees of confidence and binary belief a third binary doxastic attitude he calls 'acceptance.' He takes the term 'accept' to function in a fairly ordinary sense, similar to the sense in which we accept a proposition for the sake of argument. Acceptance is what a person is willing to assert within the context of inquiry, or alternately "a person accepts P just when she is willing to assert P in the context of inquiry; that is, just if, faced with a decision problem wherein her sole aim was to assert the truth (as it pertains to P), where her only options were to assert that P, assert that not-P, or make neither assertion, she would prefer to assert P."[171] Acceptance is an alternative to categorical belief. Whereas categorical belief may be understood as a grade of confidence above some threshold, or as a state of certainty (a grade of confidence at the maximum threshold), acceptance is supposed to be more like supposition within a context. Acceptance itself is not defined in terms of confidence rankings (above a threshold, or as a state of certainty). According to Kaplan, to define it as such would make it akin to categorical belief. On Kaplan's account, a person could rationally accept a theory (in a context. or for some purpose), but at the same time be confident that it is false.[172]

Kaplan's notion of acceptance is supposed to be a primitive cognitive state, like that of naïve belief, creditable to an agent. However, Kaplan is clear that he does not wish to replace the notion of naïve belief with acceptance.[173] A naïve belief is one that is not subject to confidence rankings. For example, an ordinary belief formed by the act of perception is a kind of pre-reflective naïve belief state. If I perceive in the ordinary way, I am forming beliefs that I

[171] Mark Kaplan, "Decision Theory as Epistemology," in *The Oxford Handbook of Epistemology*, Paul K. Moser, Ed. Oxford University Press: New York, NY (2002), pp. 451. In *Decision Theory and Philosophy* (1996), Kaplan changes his terminology from "accept" to "believe."

[172] Op. cit., p. 452.

[173] Op. cit. On his view categorical naïve belief is neither a state of confidence nor certainty, and he labels this state 'acceptance' in order to disassociate it from our ordinary conception of belief.

uncritically take to be true. For instance, I believe that I see a white piece of paper in front of me now. Such beliefs are pre-reflective in that they come to us immediately from our perceptual states. However, some naïve belief states are post-reflective, such as a belief formed by the act of “observation”. If I am observing some event in the ordinary way, I form beliefs about them, but may or may not be fully confident in the truth of these beliefs. The idea is that observation involves active attention to belief formation via a critical process, and as such is subject to confidence rankings, whereas belief about the immediately perceived is passive and uncritical.

Kaplan locates the notion within the context of inquiry, but I think inquiry should be understood more broadly to include the context of experience. For example, we might accept something for the sake of argument, but in the same way we also might accept the involuntary beliefs that we form from perception. Suppose I am walking along a sidewalk on a sunny day. I am scanning the area in front of me unreflectively in order to move down the sidewalk. As I take in the information from my perceptual processes, I accept what I perceive uncritically (at least until I encounter some anomaly that gives me reason to reflect critically on what I apparently perceive). It often happens that if given the choice of asserting some P about my surroundings, asserting some not-P about my surroundings, or asserting neither, I would prefer to assert P. This is the sense of ‘accept’ that can be applied to ordinary processes that do not involve ‘higher-order’ reasoning, but ‘lower-level’ cognitive perceptual processes.

Now suppose that the context of inquiry is more narrowly restricted to the framework of an argument. When we accept a proposition for the sake of argument, we temporarily grant the truth of the proposition within the framework of the argument, though we could have some confidence that the proposition is strictly speaking false. Borrowing Kaplan's example of the lottery paradox as an illustration,[174] we can at the same time assert that some ticket will win and have a high degree of confidence that each individual ticket will lose without running the risk of contradiction.

We can apply this more narrow sense of the context of inquiry to the example I gave above of simple, casual acts of perception, but where the context is experience and action rather than argument. When we uncritically perceive our surroundings, we grant for the sake (not of argument, but) of getting on in our lives that we are not brains in vats, that other minds exist, and so forth. In other words, in simple casual perception we accept what it is that we perceive from experience; we accept it for the sake of action. For example, if given the choice of asserting that the sun is out, asserting that the sun is not out, or neither, I prefer (for the sake of getting on with my life and the actions I need to take to do so) to assert that the sun is out, though I might have a confidence strength to some small degree that I may be dreaming. I prefer to assert that the sun is out because it is what I experience and the basis on which I act.

This narrow sense of the context of inquiry can be refined even further to a notion of the type that involves *observation*. Observation is distinct from perception in that observation is more epistemically active, requiring attention to

[174] Kaplan (2002), op. cit., p. 454.

things, both those directly perceived by the senses and those that cannot be directly perceived.[175] The added attention employed in observation involves a kind of epistemic activity requiring discernment and judgment. When we assess evidence in a scientific setting, someone (the researcher or an assistant) engages in observation of the evidence (or at least observation of instrument reading that indicate the evidence). When we observe evidence within this context of inquiry, we accept the evidence—accept it in a way that does not involve some degree of confidence we have in the evidence.

Kaplan tries to distinguish several distinct binary notions of belief. But, I think the notion Kaplan explicates is a quite ordinary notion, and that the introduction of a new term, ‘acceptance’, is of no particular help. The belief state is categorical and naive in the sense that it is not subject to ranking in terms of degrees of confidence either pre-reflectively or post-reflectively. Kaplan’s contribution is to point out that binary belief often functions relative to a context of inquiry. Contexts of inquiry vary and differ in important ways. They range from contexts like that of everyday perception to that of refined observation in an experimental setting. In each of these contexts *belief* is a binary cognitive state; given the options of asserting the propositional content of the belief, asserting the negation of the propositional content of the belief, or not asserting anything, an

[175] Bogen, Jim, "Theory and Observation in Science", *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/science-theory-observation/>. Bogen gives a historical example of “‘artificial observation’ which does not involve direct sense perception of the phenomena, but gives an indirect indication of the phenomena.” Helmholtz used a galvanometer to measure the duration of a current passing through a coil to estimate the speed of impulses traveling though a motor nerve. Helmholtz assumed that the magnitude of the deflection of the galvanometer needle was proportional to the time it took for the current to pass through the coil.

agent would prefer (among only these three options) to assert that content (sometimes even despite having only a small degree of confidence in the belief).

It seems to me that there are epistemic norms that govern the formation and possession of this kind of belief state in a fundamental, regulative way, and there are cognitive virtues and vices that affect the agent's ability to satisfy these norms. For example, when the belief state is pre-reflective (and the context of inquiry is sufficiently broad), the cognitive virtue is more akin to a skill or ability to pay proper attention. Perceptual excellence in this case just is the faculty functioning in the proper way, a way that is conducive to making the belief states that result from the exercise of the virtue that governs the relevant faculty more likely to be true than false. When the belief state is post-reflective (and the context of inquiry is narrowed), the cognitive virtue is more akin to a character trait virtue. Observational excellence perhaps involves the perceptual excellence to some extent, but the character-trait that promotes excellence, as part of the agent's cognitive makeup, is more involved. For instance, the cognitive makeup of the agent will also involve motivation as a contributing factor to observational excellence.

Where does the conception of epistemic virtue fit into the Bayesian approach to an epistemology for the sciences? More specifically, how does it fit into the probabilistic apparatus? I think that agent reliability and motivation are part of the background information we take as granted in the context. Given that an agent is epistemically reliable (employs reliable epistemic processes), he will reliably form true beliefs—perhaps only believing those claims for which he has the appropriate degree of confidence. If formation of the belief (or the acquisition of an appropriate degree of confidence) involves motivation, to the extent that the agent is properly motivated to seek the truth and has the required capacities for finding it, the agent should be reliable in forming true beliefs (or reliably holding the appropriate degree of confidence).

The idea that reliability can be fitted into the Bayesian apparatus has recently been explored by Bovens and Hartmann (2003). They construct a model of witness reliability, using degree of reliability as an endogenous propositional variable.[176] They argue that since true-positive rates and false-positive rates are well documented for medical tests, when we receive agreement on outcomes from various already well-confirmed medical tests, this does not usually influence our confidence in the reliability of the tests. However things are very different when we are considering information from sources for which the reliability is open to question. When we receive information from a set of sources, we may at first be skeptical. However, if we receive the same information from each of various independent sources, our confidence in the reliability of these sources increases. Bovens and Hartmann supply precise Bayesian models of how this should work.[177]

I think that this is an important idea for a proponent of BCT to consider. To the extent that BCT is an idealization, the difference between a partially reliable medical test and a partially reliable witness is usually not taken into

[176] Luc Bovens and Stephen Hartmann, *Bayesian Epistemology*, Clarendon Press: Oxford, UK. 2003, pp. 57-60. This model is for a single, one witness report.

[177] Ibid, p. 56.

account. Both involve evidentiary mechanisms. However, when using a medical test with well-documented and sufficiently high true positive rates and sufficiently low false positive rates, our degree of confidence in the reliability of test does not change upon receiving a series of agreeing outcomes. And that is as it should be. However, we do become more confident in the reliability of an individual witness when we receive a series of reports that can be corroborated with other reports from other witnesses. And that is also as it should be. Indeed, even in cases involving a single report this difference is apparent. Upon receiving a positive outcome for a single test with well-documented high true-positive and true-negative rates, our degree of confidence in the reliability of the test should not change. But our degree of confidence in the reliability of a witness may well change significantly when the witness provides information that we later verify as true.

Both BCT and VE are kinds of naturalism in a broad sense. The broad sense of naturalism that I have in mind is that which takes into account the natural features of a human agent, and then tries to define the notions of justification, knowledge, etc. in terms of those features. This kind of subjectivity that plays roles in BCT and in VE is only to be expected in naturalistic accounts of the attempts of human agents to justify belief and obtain knowledge.

Section 4—Virtue Epistemology meets the *Problem of the Priors*

So, how might we draw the notion of epistemic character from virtue epistemology into Bayesian epistemology to solve the *problem of the priors*? The problem of the priors is a problem for establishing constraints other than the mere rules of probability theory on what prior probabilities are assigned to hypotheses. I think that epistemic virtues may be brought to bear in reasoning constraining how prior probabilities are to be assigned. Before I delve into how this could be accomplished, I want to point out an important aspect of the problem of the priors.

In some cases the prior probabilities of hypotheses merely concern the relative frequencies of events of some sort – e.g. base rates for a disease in a population. In such cases prior probabilities are really a kind of likelihood, and are also unproblematic.[178] But the prior probabilities of most scientific hypotheses are not like this. There is no base-rate to which to appeal in assessing the prior probability of a theory of the nature of matter and energy, or for a theory about the origin of life on earth. Nevertheless, there would be no problem of assessing the values of prior probabilities if the kind of relevant background information that functions as premises for plausibility arguments for alternative hypotheses, and its strength in support of various hypotheses, were uniformly agreed to by all scientists in the relevant discipline. The problem of the priors exists precisely because such background information as provides for plausibility arguments is not uniformly recognized and agreed to, and the weight of their support for the various alternative hypotheses is not objectively or inter-subjectively agreed to by the relevant community.

It is relatively uncontroversial that individual agents hold some beliefs more firmly than others, and I think it is clear enough that these belief strengths *can be represented* in terms of probabilistic weightings on a scale from zero to

[178] I mean here that some prior probabilities really are likelihoods, Pr(E|H & K), based on frequency information for a population. For example, base rates of the incidence of HIV for various groups are these kinds of prior probabilities.

one, where the belief strengths of incompatible claims sum to the strength of their disjunction. Thus degrees of belief measured as personal probabilities make good sense, even if the numerical values are overly precise; for, it's really the ordering by strength among beliefs that matters. Furthermore, it seems to me that this subjectivist idea about belief weightings need not appeal to relative frequencies. For example, one may well believe more strongly that Alzheimer's disease is caused by reduced synthesis of the neurotransmitter acetylcholine than that it is caused by amyloid beta deposits, without basing these belief strengths on any kind of frequency information about rates at which diseases of various kinds are caused by various mechanisms. Personal probabilities represent subjective uncertainties – subjective belief strengths. Subjectivist Bayesians construe prior probabilities for hypotheses in just this way. By contrast, Bayesian likelihoods for evidence according to hypotheses are relatively objective in scientific contexts. They are taken to represent what the hypotheses *say* about the evidence. But prior probabilities are usually construed as the personal belief strengths of agents with regard to the truth of various hypotheses before the evidence is taken into account. Bayes' theorem computes the posterior probabilities of hypotheses, which represent the plausibility of hypotheses based on the evidence together with the prior plausibilities of hypotheses (as represented by prior probabilities). Thus, prior probabilities are the most subjective element of Bayes' theorem.

However, in scientific settings prior probabilities are not *merely* subjective personal probability assignments—at least they shouldn't be. In that setting all kinds of background information, especially plausibility arguments, are essential to assessing hypotheses. For example, in discussions of scientific methodology, Pierre Duhem (the physicist, philosopher of science, and anti-inductivist) often appealed to scientific "good sense", the intuitive reasoning ability of scientists, to account for how a scientific discipline may be successful in evaluating hypotheses. Among those attributes of *reasonable scientists* Duhem seems to have in mind are *good judgment* in assessing which possible hypotheses (or theories) are worthy of consideration (i.e., which ones pass the laugh test), and a fair and impartial assessment of the weight of evidence.[179]

Recently attempts have been made at reconstructing Duhem's notion of "good sense" as epistemic virtues within a version of a virtue epistemology.[180] I am sympathetic to these attempts. But as an avowed inductivist I take the notion of "good sense" to best be captured as the kinds of plausibility assessments scientist draw on in making judgments that are relevant to which hypothesis or theory is best confirmed overall. In a Bayesian confirmation theory hypotheses and theories are evaluated relative to a body of evidence *and* relative to plausibility considerations that bear on hypotheses independently of that body of evidence. The body of evidence influences the support for a hypothesis via how likely the evidence is according to that hypothesis as compared to how likely the evidence is according to competing hypotheses. Thus, the evidence impacts

[179] D. J. Stump, "Pierre Duhem's Virtue Epistemology" in *Studies in the History and Philosophy of Science*, Vol. 68 (2007), pp. 149-150.

[180] E.g. Stump, op. cit. and M. Ivanova, "Pierre Duhem's Good Sense as a Guide to Theory Choice," *Studies in History and Philosophy of Science*, Vol. 41, No. 1, March 2010, pp. 58-64.

hypothesis via ratios of likelihoods. That is the likelihoodist component of the Bayesian model of evidential support.

But there is clearly more to hypothesis evaluation than that. To see the point, consider the following kind of case. A scientist submits a research paper to a reputable journal reporting *extremely strong* experimental evidence against a new hypothesis X that no one has ever tested before. The journal refuses to publish, saying that no one would have believed X anyway – that X is an extremely implausible hypothesis – it doesn't "pass the laugh test". But isn't science supposed to be completely objective, looking only at what the evidence is? If that were true, then refuting evidence for any hypothesis should be as important as refuting evidence for any other hypothesis. The point is that scientists take plausibility considerations into account all the time. The Bayesian approach just codifies the role of such considerations. The fact that the plausibility considerations don't come from experimental evidence doesn't mean that these considerations are mere subjective whims. Scientist may (and should) have good reasons for such assessments, and they should be willing to make those considerations public, open to discussion and assessment, and reassessment. The open discussion among a community about what is reasonable and plausible is precisely the arena where the intellectual virtues should play an important role. The good epistemic character of the participants means that they are willing to engage in a fair and open-mined (epistemically virtuous) search for the truth. Real agents may only approximate this ideal. But to the extent that they do approximate it, and provided that such epistemic virtues are indeed truth-promoting, this way of employing the Bayesian logic furthers the chances that the Bayesian mechanism will produce highly confirmed hypotheses that are indeed true.

Chapter 2 explained how the Bayesian inferential mechanism can succeed here, provided the epistemic virtues employed in the service of plausibility analyses are reliable enough that they judge hypotheses that really are true to be *not too extremely implausible* prior to bring evidence to bear.

Section 5—Virtue Epistemology meets *Rational Analysis*

What does virtue epistemology have to offer the contemporary project in cognitive science known as rational analysis? If the goal is just to model how human agents extract information from an environment, then does virtue epistemology have anything to offer cognitive science? One response might be this: the way in which *rational analysis* models human behavior adds significantly to our understanding of that behavior, and the kind of Bayesian analysis employed by *rational analysis* does seem to show that some apparent cognitive mistakes are indeed 'rational' after all. So, there is no need for any additional normative standard that a virtue theory might bring to bear.

I agree that the models employed by *rational analysis* may do a good job at explaining some kinds of human behavior. Furthermore, in most everyday cases those suboptimal solutions rationalized by *rational analyses* may well be "good enough" for our needs. However, if some of the suboptimal inferential behaviors were transferred as "rational enough" to the domain of scientific inferences in real scientific contexts, scientific knowledge would be the worse for it.

Consider, for example, the kind of inferential behavior rationalized by *rational analysis* in the context of the Wason task. Recall that in those experiments most people exhibit a verification *bias*. They naturally tend to look only for confirming evidence, but fail to be disposed to look for refuting or falsifying evidence. Rational analysis does model such behavior as an optimal use of resources under the right kinds of circumstances. But that doesn't mean that a *verification bias* is always to be preferred. Indeed, were it widespread within a scientific domain it would be devastating to the attempt gain scientific knowledge. Scientific hypotheses would less often be subjected to experiment that might falsification or refutation them. As a result popular hypotheses and theories that are in fact false tend to hold on for a longer period of time. Their continued positive successes would make the search for alternatives seem much less important.

The rationalization of verification bias given by *rational analysis* may be the correct behavior (according to a decision theoretic cost-benefit analysis) in situations that only involve the "extraction of information" from an environment. And this may well help to explain the usual reliability of many of our judgments in our daily lives. But when out judgments may have much more dire consequences, the cost-benefit factors change, and getting to the truth may be more crucial.

For example, when I wake up in the morning and first open my eyes, I reliably tend to form true beliefs about my environment. The cognitive virtues that I exercise in that kind of setting are of the perceptual kind, where all I need to do is be successful is pay sufficient attention to my environment so as not to trip over the cat. In this environment, under these circumstances, by not spending too much time trying to find evidence to falsify my hypothesis about where I am and the layout of my room, I'm not only acting rationally, I am also acting in accord with epistemic virtue. For my purposes—getting to the coffee—may failure to test my environmental hypotheses is not a failure of rationality. Indeed, working too hard at finding falsifying evidence about my environment would be epistemically vicious.

However, it is worth noting that in the Wason task itself a verification bias is far from praiseworthy. In this experimental situation, given the task as described by the experimenter, failure to check for the falsifying evidence exhibits the epistemic vice of failing to look for easily accessible refuting evidence. In fact, once the mistake is explained to subjects, they invariably get it. Furthermore, when the same sort of rule is cast in less abstract terms (e.g. "if a person is drinking a beer, in then she has an ID saying she is that says is at least 21"), subjects are much more likely to look for falsifying instances (checking beer drinkers' IDs to see if the rule has not been followed).

My point is that it seems to me that the notion of rationality employed by the rational analysis program goes too far in its efforts to provide a rational explanation of the real inferential failures of people. Drawing on the notion of epistemic virtues understood in terms of skills and abilities together with the appropriate motivation component, changes how we may judge such behaviors in important ways. The intellectual virtues provide normative standards for good thinking. The usual (abstractly stated) Wason task *is* a deductive reasoning problem. People do in fact come up short in the rigorous pursuit of relevant evidence here—they fail to exhibit a virtue of good inferential practice. However, as pointed out earlier, the rigorous pursuit of relevant falsifying evidence would be a vice in a different kind of situation (getting to my morning coffee).

As with virtue ethics, the virtue, the excellence, comes with finding an action appropriate to the situation. Indeed, this is often how moral virtue is identified. For example, virtue ethics is neither rule-governed nor concerned strictly with promoting a particular optimal consequence (e.g. utility). Epistemic virtue works the same way. The virtue, the excellence, appropriate to the epistemic act is often better identified by looking to an exemplar of that excellence, and the exemplar is identified, usually, by her success at achieving the epistemic goal. Perhaps even models of ideal agents may be able to play this exemplary role in cases where it can be shown that the agent of the ideal model will tend to achieve the desired kinds of success (such as avoiding Dutch books, or converging on the truth as evidence increases).

Section 6—A Closing Remark

As I see it, the ideas discussed in this dissertation are only the beginning of a much larger project that should ultimately produce a more unified epistemology. If this investigation is headed in the right direction, then that unified view will be a virtue epistemology that is informed by Bayesian approaches to the confirmation of scientific hypotheses, by Bayesian decision theory, and by empirical research on real human cognition, as exemplified by *rational analysis*. A truly comprehensive epistemology will have to draw on all of these resources, and explore connections among them much more fully. I hope to have at least illustrated the fruitfulness of this kind of approach, and to have taken a few steps in the right direction.

Bibliography

Alston, William, "Foundationalism" in *A Companion to Epistemology*, Jonathan Dancy, Ernest Sosa, and Matthias Steup, (eds.) Wiley: Malden, MA (2010), pp. 382-385.

Anderson, John R., "Is Human Cognition Adaptive?" in *Behavioral and Brain Sciences* Vol. 14 (1991), pp. 471-517.

Anderson, John R., "The Place of Cognitive Architectures in a Rational Analysis", in *Architectures for Intelligence*, K. Van Len (ed.), Erlbaum: Hillsdale, NJ (1991), pp. 1-24.

Aristotle, *The Nichomachean Ethics*. 2nd Ed. Trans. Terence Irwin. Hackett: Indianapolis, IN (1999).

Bogen, Jim, "Theory and Observation in Science", *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/science-theory-observation/>.

Bonjour, Laurence, "Foundationalism and the External World," in *Noûs* Vol. 33, Supplement: *Philosophical Perspectives*, No. 13, Epistemology (1999), pp. 229-249.

Bonjour, Laurence, *The Structure of Empirical Knowledge*, Harvard University Press: Cambridge, MA (1985).

Bovens, Luc and Stephen Hartmann, *Bayesian Epistemology*, Clarendon Press: Oxford, UK (2003).

Chater, Nick and Mike Oaksford, "The Probabilistic Mind: Prospects for a Bayesian Cognitive Science" in *The Probabilistic Mind: Prospects for a Bayesian Cognitive Science*, Nick Chater and Mike Oaksford (eds.), Oxford University Press: New York, NY (1999), pp. 3-31.

Chater, Nick and Mike Oaksford, "Rational Analysis and Human Cognition" in *Reason and Nature: Essays in the Theory of Rationality*. J. L. Bermudez and A. Millar (eds.), Clarendon Press: Oxford, UK (2002), pp. 135-174.

de Finetti, B, *La Prevision: Ses Lois Logiques, Se Sources Subjectives* (Annales de l'Institut Henri Poincare 7 (1937), pp. 1-68. Trans. Henry Kyburg. Reprinted in Kyburg and Smokler, *Studies in Subjective Probability* Krieger: New York. (1980).

Earman, John, *Bayes or Bust? A Critical Examination of Confirmation Theory*. MIT Press: Cambridge, MA (1992).

Feldman, Richard, “Methodological Naturalism in Epistemology,” in *The Backwell Guide to Epistemology,* John Greco (ed.), Blackwell: Malden, MA (1999), pp. 170-186.

Feldman, Richard, "Naturalized Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epistemology-naturalized/>.

Feldman, Richard and Earl Conee, “Internalism Defended” in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), pp. 407-421.

Feldman, Richard and Earl Conee, “The Generality Problem for Reliabilism” in *Philosophical Studies* Vol. 89 (1998), pp. 1–29.

Fitelson, Branden and James Hawthorne, “The Wason Task(s) and the Paradox of Confirmation” in *Philosophical Perspectives*, No. 24, (2010), pp. 207-241.

Floridi, Luciano, “Logical Fallacies as Informational Shortcuts” in *Synthese* Vol. 167, No. 2 (2009), pp. 317-325.

Fumerton, Richard, "Foundationalist Theories of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/justep-foundational/>.

Garey, Michael R. and David S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman: New York, NY (1979).

Garson, James, "Connectionism", *The Stanford Encyclopedia of Philosophy* (Winter 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2010/entries/connectionism.

Gettier, Edmund, “Is Justified True Belief Knowledge?” in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), pp. 192-193.

Gigerenzer, Gerd, *Rationality for Mortals: How People Cope with Uncertainty*, Oxford University Press: New York, NY (2008).

Gigerenzer, Gerd and Peter M. Todd, *Simple Heuristics That Make Us Smart*, Oxford University Press: New York, NY (1999).

Goldman, Alvin, “What is Justified Belief?”, in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), pp. 333-347.

Greco, John, *Achieving Knowledge: A Virtue Theoretic Account of Epistemic Normativity*. Cambridge University Press: New York, NY (2010).

Greco, John, “Knowledge as Credit for True Belief,” in *Intellectual Virtue: Perspectives from Ethics and Epistemology*, Michael DePaul and Linda Zagzebksi (eds.), Clarendon Press: New York, NY (2003), pp. 111-134.

Greco, John, *Putting Skeptics in Their Place: The Nature of Skeptical Arguments and Their Role in Philosophical Inquiry*, Cambridge University Press: New York, NY (2000),

Greco, John and John Turri, "Virtue Epistemology", *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2010/entries/epistemology-virtue/.

Hacking, Ian, *An Introduction to Probability and Inductive Logic*. Cambridge University Press: New York, NY (2001).

Hajek, Alan and Stephen Hartmann, "Bayesian Epistemology" in *A Companion to Epistemology*, Dancy, J., Sosa, E., Steup, M. (eds.), Wiley: Malden, MA (2010) pp. 93-106.

Hajek, Alan, “What Conditional Probability Could Not Be”, *Synthese*, Vol. 137, No. 3, December (2003), pp. 273-323.

Hawthorne, James, "Inductive Logic", *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/win2011/entries/logic-inductive/>.

Hawthorne, James, “The Lockean Thesis and the Logic of Belief” in Franz Huber and Christoph Schmidt-Petri (eds.), *Degrees of Belief*, Synthese Library 342, 2009, Springer, pp. 49-74.

Hawthorne, John, *Knowledge and Lotteries*, Clarendon Press: New York NY (2004).

Hayden, Grant and Stephen Ellis, “Law and Economics after Behavioral Economics”, *Kansas Law Review*, Vol. 55, pp. 629-675.

Ivanova, Milena. “Pierre Duhem’s Good Sense as a Guide to Theory Choice,” *Studies in History and Philosophy of Science*, Vol. 41, No. 1, March 2010, pp. 58-64.

Kahneman, Daniel and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk", *Econometrica* No. 47, (1979), pp. 263-291.

Kaplan, Mark, *Decision Theory and Philosophy,* Cambridge: New York, NY (1996).

Kaplan, Mark, "Decision Theory as Epistemology," in *The Oxford Handbook of Epistemology*, Paul K. Moser, (ed.) Oxford University Press: New York, NY (2002), pp. 434-462.

Kim, Jaegwon, "What is 'Naturalized Epistemology?", in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell, Malden, MA (2008), pp. 538-551.

Kolmogorov, Andrey Nikolaevich, *Foundations of the Theory of Probability 2nd English Ed.*, Chelsea Publishing Company: New York, NY (1956).

Kornblith, Hilary, "Investigating Knowledge Itself" in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell: Malden, MA (2008), pp. 646-659.

Kuhn, Thomas, *The Structure of Scientific Revolutions.* University of Chicago Press: Chicago, IL (1962).

Kvanvig, Jonathan, "Coherentist Theories of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/justep-coherence/>.

Nozick, Robert, *Philosophical Explanations,* Belknap Press: Cambridge, MA (1981).

Oaksford, Mike and Nick Chater, *Bayesian Rationality: The Probabilistic Approach to Human Reasoning.* Oxford University Press: New York, NY (2007).

Pappas, George, "Internalist vs. Externalist Conceptions of Epistemic Justification", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2008/entries/justep-intext/.

Rachels, James, "Active and Passive Euthanasia" in *Contemporary Moral Problems*, James E. White (ed.), Wadsworth: Boston, MA (2012), pp. 154-163/

Riggs, Wayne, "Reliability and the Value of Knowledge" in *Philosophy and Phenomenological Research* Vol. 64, No. 1 (Jan., 2002), pp. 79-96.

Riggs, Wayne, "Why Epistemologists Are so Down on Their Luck" in *Synthese*, Vol. 158, No. 3 (2007), pp. 329-344.

Riggs, Wayne, “The Real Value of Knowing that P” in *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* Vol. 107, No. 1 (Jan., 2002), pp. 87-108.

Riggs, Wayne, “Two Problems of Easy Credit”, in *Synthese*, Vol. 169, No. 1 (2009), pp. 201-216.

Roush, Sherilyn, *Tracking Truth: Knowledge, Evidence, and Science,* Clarendon: New York, NY (2005).

Sober, Elliot, *Bayesianism: its Scope and Limits. Proceedings of the British Academy,* Vol. 113 (2002) pp. 21-38.

Sosa, Ernest, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), “Introduction: Part V” in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell: Malden, MA (2008).

Sosa, Ernest, “The Raft and the Pyramid” in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell: Malden, MA (2008), pp. 143-164.

Stump, D. J., “Pierre Duhem’s Virtue Epistemology” in *Studies in the History and Philosophy of Science*, Vol. 68 (2007), pp. 149-159.

Talbott, William, "Bayesian Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epistemology-bayesian/>.

Wason, P. and D. Shapiro, “Natural and Contrived Experience in a Reasoning Problem”, *Quarterly Journal of Experimental Psychology,* No. 23 (1971), pp. 63-71.

Wason, P. and P. N. Johnson-Laird, *Psychology of Reasoning*, Harvard University Press, Cambridge, MA (1972).

Williamson, Jon, “Objective Bayesianism with Predicate Languages” in *Synthese* Vol. 163, No. 3 (2008) pp. 341-356.

Zagzebski, Linda, “The Inescapability of Gettier Problems” in *Epistemology: An Anthology*, Earnest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath (eds.), Blackwell: Malden, MA (2008), pp. 207-212.

Zagzebksi, Linda, “The Search for the Source of the Epistemic Good”, in *Metaphilosophy,* Vol. 34, Nos. 1-2 (January 2003), pp. 12-28.

Zagzebski, Linda, *Virtues of the Mind*, Cambridge University Press: New York, NY (1996).