Title
Truth, belief, and inquiry : a new theory of knowledge

ISBN
9781321020434

Author
Fleming, Forrest Shoup.

Publication Date
2014-03-01

Peer reviewed

UNIVERSITY OF CALIFORNIA,
IRVINE

Truth, Belief, and Inquiry: A New Theory of Knowledge

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Forrest Shoup Fleming

Dissertation Committee:
Professor Sven Bernecker, Chair
Professor Jeffrey A. Barrett
Associate Professor Casey Perin

2014

DEDICATION

To my wife Anita

for her excellence and her patience,

her support and her love

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to express my deep thanks to my committee chair, Professor Sven Bernecker. Without his guidance and diligence, this dissertation would not have been possible.

I would also like to thank my committee members, Professors Jeff Barrett and Casey Perin, whose enthusiasm, expert knowledge, and generous assistance and enthusiasm for Peircian epistemology were beyond value.

In addition, I thank Professor Ermanno Bencivenga, whose insight and friendship has had a profound effect on my thought, philosophical and otherwise.

Finally, I thank the University of California, Irvine for its complete financial support in the form of a Chancellor's Fellowship, a School of Humanities Summer Dissertation Fellowship, and consistent appointment as a teaching assistant and teaching associate.

CURRICULUM VITAE

Forrest Shoup Fleming

Education

Ph.D. Philosophy - *University of California, Irvine* – 2014
Dissertation: "Truth, Belief, and Inquiry: A New Theory of Knowledge"
Dissertation Committee:
Sven Bernecker, Chair (UCI)
Jeff Barrett (UCI)
Casey Perin (UCI)

M.A. Philosophy – *University of California, Irvine* - 2010
Portfolio Papers:
"Assumptions and Cardinality: A Problem in Modal Metaphysics"
"Descartes on the Union"
"Conquest and Covenant: Hobbes on Sovereignty by Acquisition"

B.A. Philosophy – *Purdue University* - 2006
Minors: Computer Science, Political Science, English, Religious Studies

Areas of Specialization
Epistemology, Early Modern Philosophy

Areas of Competence
Philosophy of Logic, Early Analytic Philosophy, Metaphysics, Ethics

Awards and Honors
School of Humanities Summer Dissertation Fellowship – 2013
UCI Chancellor's Fellowship – 2013-2014

ABSTRACT OF THE DISSERTATION

Truth, Belief, and Inquiry: A New Theory of Knowledge

By

Forrest Shoup Fleming

Doctor of Philosophy in Philosophy

University of California, Irvine, 2014

Professor Sven Bernecker, Chair

My dissertation lies at the philosophical intersection of the American pragmatist tradition and contemporary epistemology. By treating truth, justification, and belief as matters of degree, I develop a measure of knowledge that captures all of our fundamental intuitions while providing answers to the problems of epistemic luck, skepticism, and scientific pessimism.

Traditionally, knowledge is understood as justified true belief that is not due to luck. My project follows this general outline. First, I describe the pragmatist understanding of truth first articulated by Charles Sanders Peirce in the late nineteenth century. My first chapter offers Peirce's understanding of truth as the best explanation of our intuitive understanding of what it is for a proposition to be the case and shows how we can understand Peirce's theory as compatible with contemporary theories of truth.

In my second chapter, I develop a theory of belief such that an agent believes a proposition when she acts as if that proposition were a rule governing her behavior. On this view, beliefs are theoretical entities posited to make sense of other agents' actions. Following

this account of belief, I describe what it is for a belief to be true and argue that sense of truth in which beliefs are true is best understood as an approximation of the full descriptive truth.

My third, fourth, and fifth chapters are an account of justification. Chapter 3 is a descriptive account of synchronic justification: we all reject or accept propositions in accordance with maximizing the coherence of our belief-networks. Chapters 4 and 5 articulate and then defend a new measure of diachronic justification, which is a measure of the degree to which a belief is appropriately revisable and therefore embeddable in an ongoing process of fallibilist inquiry. I develop a novel formal quantification of methodological justification and show that it gives plausible results when applied to popular cases.

My final chapter brings justification, truth, and belief together into a scalar knowledge measure. I locate my theory in ongoing epistemic inquiry, describing its conceptual advantages over rival theories as well as its ability to replicate their successes.

INTRODUCTION

Epistemology is the study of how we interact with the world. We freely talk of what we know and of what we think we know. A traditional – if disputed – analysis has it that knowledge is the complex of rational justification, truth, and belief, and while we are comfortable in saying what is true, what we believe, and our reasons for so believing, saying just what all that entails is somewhat more difficult.

Some century and a quarter ago, Charles Sanders Peirce laid the foundations for a pragmatist (or, to use his term, pragmaticist) theory of knowledge. His psychological account of belief and inquiry, when coupled his proto-Neurathian view of reality, provides fertile ground for developing a theory of knowledge. Peirce's thoroughgoing naturalism and commitment to fallibilism in inquiry yield a view of knowledge that is deeply rooted in the empirical reality of human practice.

The present work is a Peircian theory of knowledge. I have taken the early works of Charles Sanders Peirce as a starting point, and have constructed a theory of knowledge that, although quite distinct from a historically Peircian epistemology, is Peircian insofar as it accepts and builds upon this genius's insight into the atoms of epistemology. These atoms, of course, are truth, belief, justification, and inquiry. While there has been renewed interest in Peirce as a historical figure in the last decade or so, much of philosophy's contact with the pragmatist tradition has come in the form of dialogue with other pragmatist-influenced disciplines like linguistics, anthropology, cognitive science, and of course semiotics. This work is unique in that it conducts a Peircian investigation of knowledge directly and, in doing so, constructs a theory of knowledge that differs significantly from that which could be considered orthodox epistemology.

My first chapter articulates a Peircian understanding of Michael Lynch's functional truth, showing how one of the many platitudes about truth – the "end of inquiry" platitude – unites and explains the rest. Indeed, Lynch's truth functionalism is alethiology's state of the art, and it is a happy result that his view is compatible with a Peircian structure. Along the way, I make sense that which true sentences describe – *reality* – and give a brief Peircian ontology. The truth discussed in this chapter is not, though, the truth at work in the locution "justified true belief." Rather, it is the full-blown complete descriptive truth, inarticulable according to our current conceptual apparatuses and the object of diachronic inquiry.

In chapter two, I explain what Peirce's view of belief as a rule for action means for the contemporary epistemologist. I develop an account of belief similar to Dennett's intentional stance view, according to which we call the rules that a system follows "beliefs" just in case we interpret that system as an agential system that genuinely acts with intention, and these actions and intentions are guided by belief. My account of belief takes it that *all* beliefs are dispositional – occurrent belief is better understood as judgment. Although this view was heterodox just a few years ago, recent scholarship in favor of this view has gained some traction (see, for example Schwitzgebel, 2010). Once belief is well-described, I describe true belief as those beliefs whose propositional representations are satisfied by the unique model articulated from a position of ideal information, the "end of inquiry." Because very few – if any – of our contemporary sentences would have their exact semantic contents preserved at the end of inquiry, we see that the truth of a true belief is not a matter of its being part of the full-blown descriptive truth, but rather in its being recognizable as a primitive and ultimately flawed articulation of that truth. This, too, is an unorthodox position (but see Elgin, 2004).

My third, fourth, and fifth chapters are an account of justification. Chapter 3 is an account of coherence among beliefs that builds on Paul Thagard's model of belief-acquisition. Our current best cognitive science tells us that humans do form beliefs according to coherence metrics; our best models of human belief-acquisition and reasoning are coherence models. As a matter of descriptive fact, human beings maximize the coherence of our belief networks. While I take it that a belief's increasing the coherence measure of an agent's network of beliefs is necessary for knowledge (and constitutes that belief's synchronic justification), I take it that all real-world cases of belief-acquisition meet this low standard. Moreover, since all people believe coherently, coherence cannot be a full account of justification.

Chapter Four develops a normative account of diachronic methodological justification. The degree to which a belief is methodologically justified is the degree to which it is embedded in truth-seeking inquiry. I develop a novel formal quantification of methodological justification – it is the difference between degree of belief and the reliability of that belief's formation – and show that it gives plausible results when applied to popular cases. It must be stressed that the measure I develop only has normative force insofar as we accept the methodological commitments that it encodes: it seeks to tie the tenacity with which we hold a belief to the reliability of the method according to which that belief was formed. We, of course, expect that our individual beliefs are true, and that the methods by which we come to believe them are the most reliable at our disposal. Thus we cannot apply the measure of methodological justification to ourselves, as we have no external measure of reliability of our beliefs with which we could compare the degree to which we believe them. While we can look at a deviant case of belief and say where an inquirer went wrong, we are not able to do so for ourselves (though we can, of course, diagnose the error of our past selves). The practical upshot, then, is that it is generally

prudent to be open-minded in inquiry because our lack of access to the full descriptive truth gives us a tendency to overstate our own reliability.

My fifth chapter is an extended defense of the measure I develop in Chapter 4. Because the measure of methodological justification depends upon measuring the reliability of a belief-formation method, an interlocutor might object that my account falls victim to the generality problem. The generality problem with respect to belief-formation methods is that we could describe any series of actions that might constitute a belief-formation method in a variety of ways, but how we describe the method greatly impacts our evaluation of its reliability. This chapter dissolves the generality and reference-class problem as they apply to belief-formation methods by making note of the fact that there *is* a well-defined method of belief-formation – coherence with our other beliefs, as articulated in Chapter 3. Indeed, we can "read off" the reliability of these beliefs' formation method as their mean degree of truth. While the generality and reference-class problems are philosophical problems in general, they can be answered in the case of the reliability of belief-formation methods.

Finally, Chapter 6 brings the previous chapters' scalar measures – of degree of truth, degree of belief, coherence, and degree of methodological justification – together to suggest a measure of degree of knowledge. That knowledge is not a matter of degree has rarely been questioned (but see Hetherington 1999), but accepting such a view allows us to characterize the sense in which a Gettier victim really *knows*: luck undermines his justification to some degree or another depending on circumstance, but need not completely undermine his claim to knowledge.

Although I reject or avoid many of the standard assumptions of a theory of knowledge, what results from this rejection has a fairly traditional look: knowledge is still justified true belief and unreliable evidence still lowers our degree of justification. I know that 2 + 2 = 4, and that

my mug is full of coffee, and randomly guessing does not yield knowledge even when I am correct. The surface structure of my theory is not very different from what has come before, and indeed can be understood as a revision of it, as a local iteration of ongoing epistemological inquiry. Given my view of inquiry that will emerge from this document, I cannot quite say what the next iteration of my theory will look like. I can only say that I am confident that some of its contents will be preserved, and express my expectation that my understandings of knowledge-relevant truth as approximate truth, belief as habit, and justification as embeddability will be among the theoretical components so preserved. I hope that I am correct.

CHAPTER 1: TRUTH

Nearly everyone accepts that knowledge is factive – if we know *p*, then *p* must be true. From the beginnings of epistemology (see Plato's *Meno* and *Theaetetus*) to contemporary discussion (e.g. Turri, 2011; Tsohatzidis, 2012; Hannon, 2013), orthodoxy has it that we can only know something if it is true (but all orthodoxy is challenged: see Hazlett, 2010; Hazlett, 2012). I fall in line with standard epsitemology and common intuition with respect to the factivity of knowledge. Nevertheless, it will be useful to provide a brief account of just what I mean when I say that we can only know true propositions, what it is for a proposition to be true, and (later, in Chapter 2) what it is for a belief to be true. Here, I discuss full-blown truth. Luckily, much of the groundwork for such a discussion has been laid by Michael Lynch.

Lynch's *Truth as One and Many* presents an account he terms "alethic functionalism," in which truth is a single property that is variably manifest according to the domain of the relevant truth-apt proposition. In the domain of claims about dogs and cats, truth is manifest as representational correspondence. In the domain of morality, concordance[1] or something like it plays the truth-role. Truth is a functional property – it is the property, *whatever it is*, that is objective, that we can be wrong about while still being justified in our belief, is a worthy goal of investigation, and to which we should conform our beliefs. Although Lynch's analysis is largely correct, more remains to be said about the functional property itself that is variably manifest according to domain, about the function that unites the various manifestations of truth. This

[1] A belief is concordant if and only if it (i) Coheres with moral framework F at some stage of investigation, and would continue to do so through all subsequent improvements to F and (ii) F's non-moral propositions are true

property was identified well over a century ago by Charles Sanders Peirce in his pragmatist theory of truth.

By "functional property," Lynch understands truth to be such that it is ontologically unified (there is but *one* truth), but distinct in manifestation across domains of discourse (that is, there are a plurality of properties that can *make propositions true*). Truth, as a single property, is captured by four basic platitudes, or defining components of the concept of truth (p. 8-12):

1. *Objectivity*: The belief that *p* is true if, and only if, with respect to the belief that *p*, things are as they are believed to be.
2. *Warrant Independence*: Some beliefs can be true but not warranted and some can be warranted without being true
3. *Norm of Belief*: It is *prima facie* correct to believe that *p* if and only if the proposition that *p* is true.
4. *End of Inquiry*: Other things being equal, true beliefs are a worthy goal of inquiry.

In particular domains of discourse, particular properties will serve to satisfy (all) the components of truth; in the case of medium-sized dry goods, for example, correspondence between the way a belief represents the world to be and the way the world is serves nicely. In other domains, such as morality, it is unclear that e.g. moral facts (which are presumably acausal) can be truly *represented*. In such cases, perhaps the property of concordance – superwarrant within a framework of beliefs whose non-moral claims are true[2] – will suffice. It is not that "truth," for Lynch, is a definite description that picks out various properties according to domain. Rather, truth is the *single* property (which satisfies the definitional components listed above), that is *manifest* variably across domain. Truth is also not a second-order property: it is not the property of having a property that satisfies the truisms. Were that the case, the property of truth would not itself satisfy the truisms which must be satisfied for a property to be properly called *truth*.

[2] That is, manifest the property that plays the truth-role according to the claim's particular domain of discourse

In taking a Peircian line on truth, I am in effect giving a specialized reading of the End of Inquiry truism. While Lynch takes End of Inquiry to simply state that, all else being equal, having true beliefs is a good epistemic goal, I take it to have a much more important role: it unites and explains the other three platitudes. I agree with Lynch both that truth is unified and that various properties can manifest truth across various domains. The unified property of truth, though, need not be explained in abstract functionalist terms, for the single property – that which fixes belief, essentially the End of Inquiry truism – accounts for the rest, yet shows itself variably according to differing domains of discourse.

Having revealed myself early as a pragmatist about truth, it will be useful to first discuss what truth-pragmatism is not. After clearing up some common misconceptions and objections to the pragmatist understanding of truth, I will move on to a more detailed discussion of truth-pragmatism and the notions of belief, doubt, and inquiry upon which it depends.

1. What the Pragmatist Does Not Think

Reference to "pragmatic" theories of truth bring to mind the hoary bugbear of truth subjectivism. The dime-store understanding of pragmatism is that a belief is true if and only if it is useful. Of course, such a theory is absurd on its face: truth is not about what is *useful* but what is *real*. To reduce truth to mere utility is to violate the objectivity platitude about truth: A belief is true if and only if things are as they are believed to be (with respect to that belief). A proposal that the truth of a proposition depends not upon the state of the world, but rather the utility that the belief holds is simply to change the topic. If this is the sort of thing that pragmatists believe, then we are indeed a silly (if useful) sort.

To be clear, the truth-as-utility view has many obvious problems. First, it just seems *right* to say that false beliefs can be highly useful (more useful, perhaps, than true beliefs). We do not want to say that Santa Claus exists because belief in him is useful for getting our children to bed on Christmas Eve, and we do not want to say that Newtonian mechanics makes true predictions in all instances because the Newtonian system works well enough in nearly all circumstances.

Of course, the precise *nature* of "utility" above is rather vague. Even suitably refined, it is hard to see how pragmatism could be true by its own lights. If truth is simply utility, it is not clear how belief *that* truth is utility is itself useful. Such a conception of truth seems to obviate the desire to form new beliefs when our old ones are demonstrably false (but still rather useful). It seems that an alternative view of truth, perhaps as correspondence, would be far more useful.

And so the dime-store version of pragmatism fails. Though I do not doubt that some pragmatist somewhere has put forth such a theory (and some interpretations of William James's account of truth are close to this), it is not Peirce's theory, and it is not mine. The utility conception of truth, though deeply flawed, does nicely bring out that true beliefs are, in fact, useful. They are useful, though, *because* they are true (and not the other way around). It is perhaps excusable that Peirce is sometimes characterized as a utility-theorist about truth. In his 1878 "How to Make Our Ideas Clear," he characterizes truth as: "[t]he opinion which is fated to be ultimately agreed to by all who investigate" (p. 38). The object of that investigation is "the real," or reality (*Ibid.)*. Thus it might be thought that truth (and that with which it is intimately connected, reality) are dependent upon whatever (finite, human) inquirers would agree upon in some indeterminate future. The basis of truth (and reality), therefore, would be consensus. This characterization is not wholly inaccurate, but incorrectly takes agreement to be the part of

Peirce's conception of truth that does the heavy lifting. For Peirce to be plausible, we must understand not *agreement* as the lynchpin of his theory, but rather *investigation*.

It is correct to state that, for Peirce, the truth is that "which is fated to be ultimately agreed [upon]" (1878, p. 38). The reason that all investigators are fated to agree upon the truth is *that it represents the real*. Investigation itself – the scientific method of inquiry, the proper method of belief-formation – tracks reality. At the ideal limit of investigation, the fallibilist method would, eventually, hit upon the truth:

> Different minds may set out with the most antagonistic views, but the progress of investigation carries them by a force outside of themselves to one and the same conclusion. This activity of thought by which we are carried, not where we wish, but to a fore-ordained goal, is like the operation of destiny. (*Ibid*.)

What, then, is Peircian reality? In a sense, it is that which is represented by the truth. It is not, however, the system of propositions (the true propositions) that makes reality real. Rather, reality is that which is independent of both individual thinkers and thought as such – it is that "whose characters are independent of what anybody may think them to be" (*Ibid.*, p. 36). Against the charge of subjectivism, Peirce admits that the nature of the real is partially constituted by the fact that the process of scientific belief-formation – the best method of inquiry – is destined to converge upon it (*Ibid.*, p. 40). This admission does not, however, imply that the basis of reality is opinion or consensus, but rather is a steadfast declaration that the scientific method at its limit would produce (must produce) a system that wholly accounts for reality itself. At the limit of scientific inquiry, there are no unexplained phenomena, no intra-theoretical contradictions, no failure of theory to account for observation. At its limit, inquiry quantifies and systematizes all that exists. Peirce's insistence that reality is in part *constituted* indicates that *that is what it is to be real* – to admit of investigation and intersubjective agreement.

To sharpen the point, consider that which lacks reality: fiction. A work of fiction depends on an individual thinker's thinking: his thought *shapes* the fiction. Reality, on the other hand, is independent of any particular thought. It is not *shaped* by investigation, but rather determines the inevitable truth to which (scientific) investigation will lead. It is true that *being real* conceptually depends upon *admitting of investigation* – if a proposition's truth or falsity does not admit even to experimental investigation, in what sense can it be true? In what sense can it be said to represent reality if it is admitted that not just our *current* best science, but the best *possible* science, the *ideal* science that *by definition accounts for all that is real* fails to account for it? The obvious answer is that such a proposition *cannot* be said to be true. It can at best be false, and at worst nonsensical.

There is an obvious objection to Peirce's account: one might think that it counts as false or wholly nonsensical certain religious claims. Even if one takes "The god as defined by such-and-such tradition exists" to be false, to say that such a proposition is nonsense is rather unintuitive: surely we do not want to say that theologians have been senselessly jabbering for the past few millennia.

The question of religion is intimately connected to another, more pointedly philosophical objection regarding the state of abstract objects. A Platonist about, say, mathematics might hold that numbers are abstract objects that exist in a third realm beyond experience. They are inherently acausal, and as such do not admit of scientific investigation. Are we to say, then, that numbers do not exist? Is it false that two and three make five? Even more worrisome, it is obvious that experimental data regarding the nature of justice and right action is not forthcoming; does this imply that justice is not part of the real?

The solution to these worries is to remember that not all claims in a scientific theory need to be directly verifiable. If the results of investigation at the limit of inquiry must quantify over numbers, then numbers exist. That we cannot locate numbers with electron microscopes or deep-space probes need not say that they are not real, but rather that we cannot see them with electron microscopes or deep-space probes – they serve (perhaps) as a sort of precondition for our access to e.g. physical laws. If science must quantify over numbers to make sense of the world that admits of empirical investigation, then we can confidently say that numbers exist. In the case of justice, we must remember that inquiry is not limited by the scope of that which is physically observable. Although we will certainly not spot the *eidos* of the good with a telescope, justice is a topic which certainly admits of intersubjective agreement. If it really is the case (as I think it is) that normative philosophy is a truth-seeking practice, then at the End of Inquiry, inquirers will have hit upon the truth. If there are *many* ways to live the good life, such that *justice*, *good,* and *right action* are relative to a particular social structure, then knowledge of such would be available at the limit of investigation, and those systems rightly regarded as (species of the) just would be understood to be compatible with justice.

There is a further, more serious (though related) objection to Peirce's understanding: the worry of multiple realizability. At the limits of inquiry, once all the data is in, it is quite plausible to think that there might be multiple ways of accounting for that data, multiple conceptual schema that can fully articulate and account for our scientific observations. Insofar as truth is defined by inclusion in the system of propositions which fully accounts for reality, the plurality of propositional systems seems to present a problem for the common view that truth is unified.

In response to the multiple realiziability objection, it is useful to refer briefly to Tarski's semantic conception of truth. Tarski's fundamental insight was that truth is a metalinguistic predicate. Let L be the language of the end of inquiry – it contains the requisite conceptual apparatus for describing the results of all empirical evidence, as well as any necessary theoretical constructs (numbers, other minds, objects, etc.). If L is to be consistent, it cannot refer to itself – it cannot include the predicate "is true." Let M, then, be a language that contains L as well as the predicate "is true." Then we can say

T: "φ" is true if and only if ψ

where φ is a sentence of L and ψ is a copy of φ into M. Now let us consider $L_1 \ldots L_n$, each of which is a systematization of empirical data and requisite theoretical constructs. To allow us to talk about the truth of sentences in the various iterations of L, we can construct $M_1 \ldots M_n$ such that M_i contains L_i and the predicate "is true," just as in the original case. Strikingly, we now have a proliferation of truth predicates, as truth is relative to the conceptual framework (that is, model) by which we describe reality. In some instances, this is unproblematic. If L_i and L_j are intertranslatable, then there is no problem: the predicates, variables, and constants of L_i are present in L_j and are related to one another in the same way, and so the truth-predicate in M_i can be applied to sentences of L_j and *vice versa*. It is irrelevant which language we speak, because the conceptual apparatus is the same. The more serious problem occurs when we stipulate that the various iterations of L are *not* intertranslatable, that there is fundamental conceptual dissonance between them.

The pragmatist might suggest that various non-intertranslatable languages might be subsumed by a richer, more descriptive language. In this case, the language of the end of inquiry would be this superlanguage, and not any of its sublanguages. The unique model that maps the

terms and predicates of that superlanguate language to reality is the correct model, and the set of true sentences is the set of sentences satisfied by that model. An anti-pragmatist interlocutor, worried about the proliferation of truth predicates, need not worry, as it is only the metalanguage corresponding to the language that accounts for all the various systems whose truth-predicate is relevant for ascriptions of truth at the end of science.

My interlocutor, of course, would not be satisfied with such a response. It is the thrust of the objection that the world might be cognizable in a plethora of ways, but that only *one* of them can be true (after all, there is only *one* reality). This objection might be made with an example: a theory of (to borrow from Maddy) cards might speak of the deck in terms of its two colors, its four suits, its thirteen face-values, or its fifty-two individual cards (Maddy, p. 106). Each theory says something *correct* about the nature of the deck of cards. Similarly, it is reasonable to think that at the end of inquiry, there might be multiple ways of (correctly) describing the world.[3]

It is difficult for the Peircian to understand just what is at stake in the above objection, for it assumes that we can make sense of truly incommensurable conceptual vocabularies. But of course we cannot give such an example – insofar as we are able to describe the incompatible languages, they can be translated into a super-language that accounts for both. In the card example, we *know* the correct theory of cards: it accounts for the deck's colors, suits, pips, faces, and individual cards as well as its unified nature as *a deck*. *All* objective properties are ascribed to the cards; any properties that do *not* admit of inquiry and the settling of belief are non-existent, fictional, not part of the real world (an analogy might be the value of a particular set of cards if

[3] Maddy's point is that the fact that there are multiple ways of describing the world need not indicate that the world has no nature of its own, that the "joints" of nature are given to it by human cognizers. Although I am using her example to illustrate an objection to a pragmatic understanding of truth, one must not attribute this objection to Maddy.

interpreted as a hand of poker). It is precisely those theoretical constructs that are *constant* among the various iterations of *L* (numbers and pips and colors suits) that we can say are in fact mind-independent, do in fact exist independent of what anyone thinks about them. Those predicates and variables and constants that *vary* from language to language, to which the metalinguistic truth-predicate does *not* always apply, are precisely those features of the world that do not exist. They are the features of the world that are wholly determined by cognizing agents. They are fictional.

The anti-pragmatist could, of course, press this point. He could make reference to past competition among theories, such as the equal empirical adequacy of complex Ptolemaic theories of planetary motion and early heliocentric systems. But what he insists upon is not the inadequacy of pragmatic conceptions of truth, but rather a thoroughgoing conflation of reality and fiction. The pragmatist does not hold that current scientific methodology reflects the methodology in place at the end of inquiry; the only constraint upon ideal scientific practice is that it is falsifiable and tracks reality. To insist that our best scientific theories will involve mind-dependent entities is to insist that reality is not amenable to scientific investigation, to insist that inquiry as such simply *cannot* account for the world. It is, in a way, to insist that fiction itself is necessary to a complete systematization of the world.

It is important to note that the anti-pragmatist is here asserting far more than that the conceptual apparatus in use at the end of inquiry might need to be extended beyond directly-verifiable sentences. The pragmatist fully agrees with *this* claim, that perhaps the natural numbers or consciousness or justice or some particular subatomic force must be posited even though it cannot be directly confirmed. Indeed, it is precisely in this way that metaphysics vindicates the natural sciences, logic, and so on for the Peircian: metaphysics posits and accepts

those entities that the physical sciences need, but cannot directly confirm. The pragmatist considers these indispensable theoretical entities as *part* of the theory – as of the same ontological status as bricks and dogs and other minds. What the anti-pragmatist asserts is that the pragmatist's view of the end of inquiry, *including* the necessary non-directly-verifiable theoretical entities (which are just as real for the pragmatist as the theoretical entities of dogs, mountains, and other minds) is *still* lacking, that the language still requires enrichment.

And perhaps the anti-pragmatist is correct in his way. Perhaps even with all the data in, humans would not be satisfied. Perhaps even with a full evolutionary history of the universe and complete laws governing everything from fundamental particles to human social interaction (be they deterministic or not), we would still find ourselves in Peirce's irritating state of doubt, and would tell ourselves stories to eliminate this irritation. But this is not a claim about reality, about the world as it is. It is a claim about human psychology, and it is a quite plausible claim. Of course, the claim reflects a certain sort of pessimism with regards to the condition of humanity – it states that we are so constituted that what amounts to a *complete* systematization of the world would still not be enough for us. But if this pessimism towards inquiry and human nature is correct, it merely indicates that humans long for something above and beyond that which really exists. If the state of humanity is so impoverished, it is not the province of a theory of truth to salve it.

The anti-pragmatist might object here. His worry is not that the world cannot be fully described, but rather that we might fully describe it in *different ways*. There might be multiple conceptual apparatuses that account for all the same data, but do so according to radically different ontological claims and properties. We would then have sets of true sentences according to each system, and it might not be clear at all which is the "correct" systematization of reality.

This seems to be the case with regards to ethical concepts like justice: if justice primarily involves a social contract agreed to without regard to particular non-moral properties of agents, then perhaps Rawls (or something like his theory of justice) is correct. If it is fundamental to justice that one's property is an extension of one's self, then justice might be seen to be something like the notion articulated by Locke and more fully developed by Nozick. The objection states that, at the end of inquiry, we might not *know* whether we should be Rawlsians or Lockeans about justice – the systems could be fully articulate and coherent, but incompatible, theories of justice.

If the above case is right, then the pragmatist is forced to say that there *is no* unique "correct" description of reality (moral or otherwise). Of course, many *incorrect* systems are ruled out at the end of inquiry, namely those that do not accurately predict observation, fail to account for some aspect of the world, contradict themselves, and so on. However, if the universe (or justice) can (like the deck of cards) be described either in terms of *F*s or in terms of *G*s, but (unlike the deck of cards) cannot be described in a system that accounts for both *F*s *and G*s, then there is simply no truth of the matter about whether we should speak about the universe in terms of *F*s or in terms of *G*s, whether we should be Rawlsians or Lockeans. If we conceptualize the universe in one way we use *F*s, and if we conceptualize it in another way we must use *G*s. In instances wherein the conceptual systems coincide, we can say that they carve nature at the joints. In the instances where they diverge, we must simply say that nature has no joints (or alternatively, is jointed everywhere).

It is clear that Peirce himself, as a sort of essentialist in an Aristotelian vein, takes the world to have a definite structure that we uncover through observation and experimentation. Later philosophers like Sartre have told us that differentiation is something that consciousness

brings to the world, that matter itself exists without meaning or structure and that its apparent structure is in fact imparted to it by consciousness. Which way we go on this question seems to be more a matter of taste and philosophical inertia than of factual importance. The Sartrean and the Aristotelian can agree that the existence of a single, discoverable structure to the world is a regulative ideal that shapes our inquiry. Whenever we believe a proposition or doubt something, it is because we expect that there is a way that the world *is*, and that our judgment as to that fact was correct or incorrect as the case may be. The further question – whether there *really is* a way the world is, or if that "way" is an artifact of the way that humans perceive, conceptualize, and so on, is less important. Indeed, it is difficult to see how a Peircian could worry too much about its answer: even if the structure of the world comes from our consciousness, there is no way to interpret the world apart from that structure, and so it seems impossible to articulate how the answer to the question could change our experience of the world. That is, the answer to the question seems particularly *immune* to inquiry, and so it is not a genuine question of reality at all.

2. Doubt, Inquiry, and Belief

In the preceding section, I transition without comment from that which is *true* to that which eliminates doubt. For Peirce, they are very nearly one and the same. Peircian doubt is that uncomfortable mental state in which we feel unprepared to act, as we recognize that the result of our action will likely be frustrated. Belief is the elimination of this discomfort, and is the comfortable state wherein we feel that we are ready for action, that our action will have the consequences we intend. In a state of doubt, we are driven to action to "fix," or settle on, our

belief, to achieve that comfortable state that conditions us to behave in a particular way when the occasion demands (Peirce, 1877, p. 10).

Peirce calls the struggle to move from doubt to belief "inquiry." He considers the psychological irritation of doubt, and the corresponding comfort of belief, to be the only motives to inquiry. We want our beliefs to be such that when we act on their basis, our action results in the desired outcome. Any belief whose formation process does not help to ensure this connection between action and outcome (or which results in an unexpected failure when used as a basis for action) is transformed into doubt; inquiry ends when the doubt is settled (that is, when we take our belief to be an appropriate basis for action). Thus inquiry is wholly psychological, bookended by doubt and belief. While we generally state that we desire not just *beliefs* but *true beliefs*, this does not appear to be the case in practice: "The most that can be maintained is, that we seek for a belief that we shall *think* to be true. But we think each one of our beliefs to be true, and, indeed, it is mere tautology to say so" (Peirce, 1877, p. 10-11). As such, the end of local inquiry is not a set of true beliefs but a set of beliefs that one takes to be true. The sole end of inquiry is not truth, but the settlement of opinion. The set of true sentences both end inquiry by transforming doubt into belief, and obviate the need for further inquiry because they are indeed true. Each of these parts is essential to understanding the nature of truth.

The "end of inquiry" in a global – as opposed to local – sense is that state in which all doubt has been removed from all inquirers by a methodology that they consider to be appropriately truth-tracking. In light of Peirce's conception of doubt, we can formulate and respond to a more pointed objection. The objection is this: at the end of inquiry (in a global sense), with all the data in, there might still be gaps in that for which our theory can account. There will still be sentences that seem amenable to truth-ascription, yet are not included in the

set of sentences "fated to be ultimately agreed to by all who investigate" (Peirce, 1878, p. 38). The existence of objects outside our light-cone will not be included, and yet it certainly *seems* that there should be particular facts regarding objects outside our light-cone even if we cannot have evidence for or against them.

This objection is in a sense unanswerable. It posits an "end of inquiry" in which all inquirers' doubts are settled – there is no question with regards to objects outside our light-cone, to make use of the example – and yet by hypothesis, there is no sentence of the form "There exists (or does not exist) an asteroid with such-and-such properties at such-and-such location." The objection highlights, I think, two issues. The first is the fundamental metaphysical issue that Peircian reality ends at precisely that point wherein it is impossible for inquirers as such to retrieve data. The second is the epistemological issue of unknowable facts, and the implication that the end of inquiry is unreachable.

In response to the first issue, it is sufficient to note that it is not *human beings* that are required at the end of inquiry, but rather inquirers as such. It is not the possibility of knowing *for humans*, but rather the possibility of knowing *at all* which is exhausted at the end of inquiry. The end of inquiry is, by definition, the state wherein *all* beliefs are fixed. For beliefs to be fixed in the long term, they must be founded in the external world – they must be about objective reality (otherwise, inquirers would still doubt). Reality is that which is amenable to investigation. It is true that Peircian reality ends at the point that it cannot be observed. It is not, however, *practical* observation by human beings (or alien creatures, or what-have-you) that is at issue, nor it is *observability according to the laws of our current best science,* but rather *observability in principle*. Peirce's definition of reality does require the presupposition that *all* reality is observable in principle – that it is the fundamental nature of existence that existent

objects should be capable of investigation either by direct empirical testing or by position in a theoretical-conceptual framework. The objection that, at the end of inquiry, we would not have reason to believe one way or the other about some particular facet of reality is the objection that the end of inquiry is not the end of inquiry, that we (in the sense of inquirers, not in the sense of human beings) will still have doubts after all. It is the statement that there is *something* which is accountable to reason in principle, but not in actuality. The solution to the objection, then, is clear: the state so described is simply not the end of inquiry properly understood.

The first issue flows easily into the second – the epistemic position that not all facts are knowable. For certain understandings of possibility, this is correct. It is simply not the case that inquirers, human or otherwise, will ever be able to account for the totality of reality, for the objects outside our (collective) light-cone and so on. This presents no objection to a Peircian system of truth, however. The pragmatist conception of truth does not include the statement that we (that is, inquirers) ever *will* reach the end of inquiry, that we will ever have epistemic access to all the facts (indeed, even if we *were* at the end of inquiry, we would not know that we were). The position, rather, is that what it *is* for something to be real is for it to exist independently of what anyone thinks of it, and that the full descriptive *truth* is the description of this reality that soothes doubt and replaces it with belief. The end of inquiry is not a point in history, but an epistemic ideal – it is the state of complete epistemic access to the external world. That such a condition will never obtain (and, on most understandings of possibility more restrictive than logical possibility, *can* never obtain) is irrelevant. Peirce explicitly answers this question in a beautiful passage from "How to Make Our Ideas Clear":

"Full many a gem of purest ray serene / The dark, unfathomed caves of ocean bear; / Full many a flower is born to blush unseen, / And waste its sweetness on the desert air." … Do these things not really exist because they are hopelessly beyond the reach of our knowledge? And then, after

the universe is dead … will not the shock of atoms continue though there will be no mind to know it? To this I reply that … it is unphilosophical to suppose that, with regard to any given question … investigation would not bring forth a solution of it, if it were carried far enough. …Who can be sure of what we shall not know in a few hundred years? Who can guess what would be the result of continuing the pursuit of science for ten thousand years, with the activity of the last hundred? And if it were to go on for a million, or a billion, or any number of years you please, how is it possible to say that there is any question which might not ultimately be solved? (p. 39-40)

Peirce is clearly a scientific optimist. The most obvious articulation of the question with regards to the relationship between epistemic access and metaphysics is the question of the existence objects beyond our light cone. Our current best science tells us that interaction with such objects (if indeed they exist) is impossible. Peirce would instead remind us first that our current best science will be supplanted, and it is foolish to attempt to predict conceptual revolution. He would further caution us against mistaking any historical point as the end of inquiry. The end of inquiry is an ideal state of omniscience, not a point in human history. Finally, he would remind us that insofar "object beyond our light cone" is a dressed-up phrase that has less to do with light cones and more with an attempt to articulate a more respectable version of "object whose observation is logically impossible," the phrase is empty.

Just as the proper view of Peirce's end of inquiry evaporates worries about unknowable facts, worries of multiple realizability similarly dissipate. At the end of inquiry, it would be a settled matter whether we should speak of the universe in terms of *F*s or *G*s. If *F*-theory and *G*-theory are equally predictive, coherent, and so on, then the inquirer would conclude that whether we speak of *F*s or *G*s (and all the theoretical baggage included either way) is simply a matter of perspective (or choice as to which conceptual-linguistic system to adopt). And at the end of inquiry, recognition of this fact would not occasion doubt as to whether the universe is *really* made up of *F*s or *really* made up of *G*s. If inquiry is over – if belief is fixed – then the *F*s-or-*G*s

question is by definition of "end of inquiry" not an occasion for doubt, and does not pose a question to be settled by the external world. This is to say that, if the universe is equally well-described in *F*-terms or *G*-terms and *F*-terms are incommensurable with *G*-terms and there can be no broader theory that incorporates both, inquirers would necessarily *recognize* that the *F*s-or-*G*s question is merely a question that is left up to choice, or taste, or perspective. If the decision is to not be an occasion for doubt (as it must not be, since the worry must occur at the end of inquiry to be relevant), then it seems that the *F*s-or-*G*s question must *really be* a question of choice, or taste, or perspective. To pose the question would be similar to asking which inertial frame is "correct" with regards to general relativity. As with inertial frames, the *F*s-or-*G*s question just doesn't matter, but once we have decided one way or the other, the rest of the conceptual framework follows.

3. Unifying the Platitudes

Peircian truth consists of two important claims. First, the set of true sentences fully describes reality. Second, we cease inquiry when we have beliefs that we take to be true. If our actions unexpectedly result in failure, or we think that actions based on our beliefs might result in failure due to the belief's being based not in reality but in some other contingent factor (such as custom, bias, or unjustified assumption), we doubt again and begin inquiry. The end of inquiry, wherein "[n]o modification of the point of view taken, no selection of other facts for study, no natural bent of mind even, can enable a man to escape the predestinate opinion," is the truth, which accounts for all that is real (Peirce, 1878, p. 38). How, though, do Peirce's

pragmatism and psychological account of belief, doubt, and action count as a theory of *truth*, how does Peircian truth account for Lynch's truisms? Let us treat them in order.

Objectivity is perhaps the most intuitively important truism. It ensures that there is a genuine connection between that which is *true* - that is, a proposition or a sentence - and that against which that which is true is measured. The explanation for Objectivity in a Peircian system is the truth-seeking nature of inquiry. It is, for Peirce, evident that humans default to the scientific method of inquiry. We doubt beliefs in response to both contradictory evidence and evidence that our now-doubted belief was not an appropriate basis for action. In the long term inquiry is self-correcting. While any individual might take any belief to be true, the opinion of inquirers as such is destined to converge upon the truth simply because the truth *really does* depict reality, and it *really is* an appropriate basis for action. It's a simple matter of fact that true beliefs are better grounds of action than false beliefs. Our psychological aversion to having our desires frustrated will lead us to true opinion.

Warrant Independence requires a bit more nuanced treatment than Objectivity. At first blush, it might appear that Warrant Independence directly contradicts Peirce's account of truth. After all, one crucial aspect of Peircian truth is that it that which is destined to be agreed upon at the end of inquiry. At the end of inquiry, it seems that we have all and only those beliefs that are warranted by (the complete set of) observations, and those beliefs are all and only the true beliefs. Thus the disconnect between the two is paradoxical.

The paradox is dissolved when we understand Warrant Independence to state that *any particular agent's* belief may be warranted and false, and that *any particular agent's belief* may be unwarranted but true. All that Warrant Independence says, then, is that good methods of belief-formation can lead us astray, and that bad methods of belief-formation can get lucky and

land upon the truth. This fact need not fly in the face of Peirce's understanding of truth. When we recognize that the end of inquiry is not a historical point in time but rather an ideal, it is obvious that there simply *are no* (former-) inquirers at the end of inquiry. "End of inquiry" is not merely the point at which we hit upon the truth, but rather is a way of speaking *about* the truth.[4] Truth ends inquiry because true beliefs are perfectly suited to form the basis for successful action, and actions based on true beliefs are less vulnerable to frustration. That is, there is no occasion for doubt – no contradictory evidence, no unsuccessful action – when we have all and only true beliefs.

For particular agents, warranted false beliefs and unwarranted true beliefs are both byproducts of an incomplete grasp of the truth, of not believing all and only true propositions. In the case of warranted false beliefs, there exists some set of true propositions that, if believed, would occasion doubt in the epistemic agent, thus leading to a reevaluation of her mistaken belief. In the case of unwarranted true beliefs, there is some set of true propositions the believing of which constitute an agent's warrant. In either case, the agent's belief would be both warranted and true if she knew all the true propositions. That she does not – and that in all likelihood human beings cannot – know everything does not tell against Peirce's understanding of truth.

The cases of End of Inquiry and Norm of Belief are, as stated, incomplete. Norm of Belief states that it is *prima facie* correct to believe *p* if and only if *p* is true. Given that any old belief we take to be true (that is, *any* belief) serves to dispel doubt, why should the true beliefs be privileged? In what sense is truth a normative constraint on belief formation? Simply put, true

---

[4] It is comparable to Kant's Kingdom of Ends in this way.

beliefs are *better* than false beliefs not in general, but in obviating further occasion for doubt. While we can certainly come to doubt true beliefs, a belief that is true will never be directly disconfirmed. Even if the true belief is unwarranted and its negation warranted, false beliefs are easily susceptible to disconfirmation in the face of further data. If one's project is to minimize doubt, then one could hardly do better than to cultivate true beliefs. Of course, particular conditions (a network of connected false beliefs, for example) might obtain that make a false belief a better bet than a true one. But such particular conditions obviously defy the *prima facie* qualification of Norm of Belief. At first glance, a true belief is less susceptible than a false belief to being the basis of a failed action or direct disconfirmation. And so at first glance, it is correct to believe p if and only if p is true. In particular circumstances, it might be the case that all available evidence points to $\sim p$ and so the rational thing to do is to fix belief on $\sim p$. But such circumstances are precisely what the qualification of "*prima facie*" allows.

Similarly, End of Inquiry states that true beliefs are a worthy goal of inquiry. For Peirce, it is not true belief that is the end of inquiry in a particular agent, but rather belief as such. It just so happens that *true beliefs* are less susceptible to future doubt than false beliefs, and so over time many inquirers working from many different starting points will converge upon the truth. It is certainly the case that true beliefs are worthy goals of inquiry, but this is not because truth itself has normative weight. True beliefs are simply better bases for action than false beliefs, because acting on a true belief is more likely to result in the realization of our desired outcome and the avoidance of doubt.

It is striking that pragmatic truth, which is oftentimes seen as a sort of relativism where "truth" loses all recognizable meaning, so easily captures our fundamental constraints upon a theory of truth. At the heart of Lynch's collection of platitudes is the core of Peirce's system,

that which makes pragmatism about truth particularly pragmatic: *true beliefs are useful*. Truth is what a Rawlsian might call a "primary good" – that which a "rational man wants whatever else he wants" (Rawls, p. 79). No matter your project, having true beliefs is better for it than having false beliefs.[5] Whatever your desires, acting on the basis of true beliefs is a better way of realizing them than acting on the basis of false beliefs. That is simply part of what it *is* for a proposition to be true. The truism of Objectivity explains *why* true beliefs are better than false beliefs – they represent the external mind-independent world. Warrant Independence assures us that particular agents can go wrong when they do not have the facts, although Lynch's original formulation must be understood to *only* apply to individual inquirers, since the set of true beliefs is identical to the set of warranted beliefs at the end of (global) inquiry. Norm of Belief and End of Inquiry are elaborations upon Objectivity, simply stating *that* true beliefs are better bases for action than false beliefs. The standard objection to truth pragmatism – that a false belief can be very useful, and a true belief useless - is true but toothless. False beliefs are always useful *for* something or other in particular. True beliefs are useful for nothing in particular; they are useful in general and for all ends. Truth is a primary good, and whoever has a true belief is (epistemically) better off for it.

Truth is a single property – the property of dispelling doubt and fixing belief in inquirers. There are many mechanisms that occasion doubt in inquirers – unexpected outcomes of actions, unexpected observational results, and contradiction are but a few. Michael Lynch's alethic functionalism, wherein truth is a single functional property that is differently manifest in various domains, is largely correct. Truth fixes belief via observation and frustrated action in the case of

---

[5] This assumes that your project isn't having lots of false beliefs or as few true beliefs as possible. I am skeptical that such a project is possible.

medium-sized dry goods, via contradiction and logical consequence in mathematics, and via concordance in morality. Lynch's identified truisms do not merely identify functions for truth to fulfill across various domains, but point the way to a single, robust property: belief fixation beyond reproach. As doubt is occasioned in various ways, so belief must be fixed according to various methods. By marrying Lynch's functional account of truth with Peirce's pragmatism, we can see that there *is* in fact a unifying property that all manifestations of truth share: the fixation of belief, the end of inquiry.

CHAPTER 2: BELIEF

In the previous chapter, I articulated a theory of truth that is ultimately relative to rational inquirers. Truth fixes belief in inquirers, and it so does because it is about reality. Being about reality, truth is a better basis for action, belief-formation, making plans, and so on than anything else. Were an inquirer's beliefs false, she would eventually be frustrated in her actions, surprised by her observations, notice a contradiction in her beliefs, or similar. Such frustration, surprise, and awareness of contradiction occasion doubt in the inquirer, and she struggles to return to the comfortable psychological state of belief. Until *all* her beliefs are true, an inquirer will have occasion to doubt; when all beliefs are fixed, she has reached the end of inquiry. This chapter will explicate my understanding of belief in general, and of *true* belief in specific.

1. Peircian Foundations

If truth is a functional property, and that function is the fixation of belief, it will be useful to elaborate on the understanding of belief in use. My account of belief, like my account of truth, is inspired by Peirce: belief is the psychological state wherein an agent feels that her action will lead to her predicted outcome. She feels prepared to judge and to act. If she receives evidence that her action might not lead to its expected outcome – perhaps she learns that her belief is better explained by cognitive bias than by conforming with facts external to her mind – her pleasant state of belief is destroyed, and doubt arises in its stead. To be in a state of belief is to have a rule for action, and to be in a state of doubt is to lack this rule.

In the cognitive sciences, the two most prominent views with respect to attributing mental states like belief, desire, disgust, and so on (in the literature, "mindreading") to others are theory

theory (TT) and simulation theory (ST). In TT, we attribute beliefs, desires, and so on to others as part of a theory that we represent in our minds (or brains). Social action is thus the application of an empirically testable theory to other agents, in which we posit an agent that behaves more or less rationally and the particular mental states of beliefs and desires. For example, if I believe that Ralph desires coffee and that he believes that there is fresh coffee in the kitchen, I can predict his decision to head towards the kitchen with something like the following process:

1. Ralph desires coffee
2. Ralph believes that there is coffee in the kitchen
3. In the absence of competing concerns, when an agent *S* desires *x* and believes that action *a* will yield the satisfaction of his or her desire, *S* will *a*.
4. There are no competing concerns for Ralph
5. Therefore, Ralph will decide to go to the kitchen, and will indeed head to the kitchen

My prediction that Ralph will go to the kitchen (and my attribution of the decision *go to the kitchen* to him) is based on a particular conceptual framework – the framework that includes *beliefs*, *desires*, *actions* and so on and specifies their interrelations. The development of mindreading, according to TT, just *is* the working out of this conceptual framework through trial and error much in the same way we establish folk physics and the corresponding implicit concepts of occlusion, solidity, containment, and so on in early childhood (see Gopnik & Wellman, 1992; Gopnik, 1996; Gopnik & Meltzoff, 1997).

Simulation theory, on the other hand, holds that we need no particular conceptual apparatus to attribute mental states to others because we simulate their thought processes ourselves. We start from another's perspective (insofar as we can adopt it) and reason as we would so do in a similar situation. For example, if believe that Ralph desires coffee, and I believe that he believes that there is fresh coffee in the kitchen, I will come to predict his actions by simulating what I would do if I had his beliefs. Accordingly, I imagine myself with a desire

for coffee and the belief that it is in the kitchen, and I use my own decision-making capabilities to decide that I will get a cup of coffee (though, of course, I do not act on this decision). I then attribute to Ralph the decision to get a cup of coffee, and predict that he will indeed head to the kitchen.

In ST, the mindreader requires no particular conceptual apparatus, and need not possess any ability to articulate movements from one mental state to another. Where a TT-mindreader must have at least an implicit understanding of how belief and desire yield action, the ST-mindreader simply inputs the relevant beliefs and desires into her *own* decision-making process and then – with no understanding of how the belief and desire create action or intention or anything else – comes to attribute decision and action to an agent based on what the mindreader would do (Gordon, 1986; Gallese & Goldman, 1998; Vogeley et. al., 2001).

Of course, this is not all there is to say about ST and TT. Indeed, TT widely acknowledges a range of position ranging from taking the relevant theory of mind to be implicit and not explicit (Martin, 1994) to being "contained" in a physical brain module that performs the relevant inferences (Scholl & Leslie, 1999). The TT theorist simply *cannot* hold that all human beings have conscious access to the same conceptual framework of belief, desire, and so on simply because the empirical evidence says the exact opposite: the relationships between beliefs, desires, and so on are incredibly dependent upon cultural and linguistic folkways, and these folkways strongly shape an agent's theory of mind (or, in fairness to ST, how their simulation runs). With the introduction of implicit theory and modular delegation of mindreading tasks, TT and ST begin to look strikingly similar. We already know that our brains possess mirror neurons that activate both when we perform particular actions and when we observe a sign that others are performing those actions. There is no doubt whatsoever that there are physical structures in our

brains that simulate (at least some of) others' mental states as if they were our own. Once we admit that the implicit theory of TT might be controlled and (at least in some cases) used without conscious conceptual understanding of the relevant theoretical constructs and constraints, we have two accounts with identical start and end points mediated by a black box of calculation.

The difference, then, lies in how each theory of mindreading explains what happens in the black box. ST says that we do the processing from a first-person point of view, while TT says that we do the processing from a third-person point of view according to relevant rules. Importantly, there is no consensus among scientists as to whether the existence of mirror neurons better supports ST or TT (see e.g. Vogeley et al, 2001; Gallese and Goldman, 1998; Schulkin, 2000; Goldman and Gallese, 2000). This seems like a significant difference, but it is not crucial to my present project of understanding belief itself.

Consider how we could develop a theory of mind in accordance with TT. The most plausible story is that we recognize that *we* have beliefs, desires, and so on and that we behave in certain ways due to their constant conjunction. We generalize to violable but generally-correct rules of thought and action. At some point, we notice that there are other things out there that bear a striking resemblance to us, and we apply the same rules to predict and explain their actions, decisions, and so on. Importantly, it is *our own minds* that are a model for the rules that make up the structure of TT's model of the mind. Thus TT says that we extract rules from our own thought processes and apply them to others, while ST says that the generalization step need not happen, since we just simulate the decision process ourselves. When we take seriously the common TT-view that the theory is implicit and not necessarily understood or articulable by the mindreader, or the more extreme view that a module takes care of this work for us, the divide between simulating and applying our own rules (of which we are not aware, thanks to their

implicitness) begins to collapse. This collapse is further heightened by the current popularity of ST-TT hybrid theories (see Mitchell, 2005; Saxe, 2005a; Saxe, 2005b; Saxe, 2009), which state that *sometimes* we (perhaps implicitly) apply rules that are generated from outside data (including the operations of our own minds) and *sometimes* we take a shortcut and offload the calculations to our own decision-making faculties.

In what follows, I will largely use language that is most conversant with the TT side of things. I do not intend this to make a claim as to what happens in the black box of neurons that constitutes the human mind, for that is either an empirical question to be settled by the scientists (of which class I am not a member) or there is no fact of the matter (and thus there is no sense in worrying too much). Rather, I will discuss what competing theories of mindreading take for granted – that we do in fact attribute beliefs, desires, and mindedness to others, and that (by some process or other, be it simulated, conceptual, a hybrid, or more likely something as of yet undreamed by our current best scientific theories) we do in fact attribute intentionality and actions to them. *This* process – the attribution of the primitive states that are preconditions to the cogency of the mindreading debate – is one that I am comfortable speaking of in terms of a theory. Indeed, I want to say what it *is* to attribute a belief to an agent and what it *is* to say that she performed a certain action. Whether the intervening steps are implicit, explicit, or simulated in whole or in part by physical structures that obviate the need for any conceptual framework is a separate question all together. In what follows, I articulate a conceptual understanding of what beliefs are, but make no claim that the bearer of beliefs has *any* understanding of beliefs, their nature, theoretical revision, or anything else. The framework I present is to be utilized by an external observer, and I make no predictions or claims about a believing subject's access to it.

2. The ABI Model

It is important to recall that beliefs, like other minds and dogs and mountains, are theoretical entities: their ontological status is posited because experience demands an object in which certain properties inhere. Just as we posit the object *dog* to make the best sense of dog-like data, we posit that agents have particular *beliefs* to make sense of belief-like data. "Belief-like" data is simply the output of a system whose operations are best explained – and whose future outputs are best predicted – by a tripartite intention/belief/action model. Any system that is best explained by the tripartite belief model is a belief system (not to be confused with a system or network *of* beliefs) (for a similar view, see Dennett, 1987). The three states of intention/belief/action are defined in terms of the belief-system itself and in terms of each other – they are the atoms of the system.

An *action* is any output of a system that is directed towards a purpose and guided by one or more rules by which the action is expected to achieve its purpose. A *belief* is a general rule by which particular outputs of the system can be expected by that system to achieve their purposes. An *intention* is the expected result the system's output, determined by general rules. More succinctly, an *action* is an output of a system that aims towards an *intention* and is guided by *belief.*[6]

Any system that is best explained by the action-belief-intention (ABI) model is a belief system.[7] In approaching a set of observations as a belief system, we posit an entity in which the

---

[6] One might notice that I class anything that is an "expected outcome" as part of an action's intention. If one wishes to define action and intention in some other way such that the person who knowingly kills another by pulling a lever doesn't *intend* to kill them but merely *intends* to serve justice, or play a lever-game, or what-have-you, this is acceptable. Such a discussion is quite beside the point at hand.

[7] The criteria of comparison should be those applied to any other theory. Popular candidates (in addition to empirical adequacy) are parsimony, simplicity, unity, depth, elegance, explanatory power and so on. Because I do

intentions and beliefs inhere and that is both the owner and cause of that system's intentional outputs (that is, its actions). We call this entity a *mind*, and we call anything that has a mind an *agent*. Agents have minds; minds have intentions and beliefs and they perform actions.

The most obvious object to which the belief-system interpretation applies is ourselves. We have enormous evidence (significantly, if not solely, from introspection) that we act for particular purposes (that is, with *intention*), and that we act according to rules that predict the success or failure of particular actions. We generally find it plausible that we *really do* author actions, and that our actions *really are* based upon our beliefs about the world. Indeed, interpreting oneself as a minded agent is inviting to the extent that we often confuses our belief that *p* with an action expressing our assent to *p* – we sometimes confuse our thought "*p*" with our belief that the world is such that *p*. When I have a desire for coffee, I intend to pick up my mug and bring it to my lips. I believe that the mug weighs such-and-such when it is full of coffee, that coffee is hot, and so on, and the veracity of these beliefs yield an action that conforms with my intention. If I had false beliefs (perhaps that my mug is much heavier than it is) my actions might be frustrated (perhaps I use too much force in picking it up, thus spilling my drink). Throughout this process, I have access (though surely not perfect or infallible access) to the intentions of my actions and the beliefs that guide them.

---

not expect that what counts as a good theory at the end of inquiry would be identical to what counts as a good theory by today's standards, I do not wish to argue forcefully for any particular set of theoretical virtues. It is worth noting, though, that last chapter's understanding of reality requires that a good theory is empirically adequate and parsimonious. It must be the former because complete empirical adequacy by definition accounts for all the data, and accounting for all the data is part of the definition of "end of inquiry." Similarly, the conceptual framework at the end of inquiry must be parsimonious simply because the definition of "end of inquiry" includes that all the data is in, and that anything for which the final theory doesn't account simply *doesn't exist*. Since the theory is maximally empirically adequate and by definition anything that isn't included in the theory doesn't exist, the theory must be parsimonious as well. Of course, just what counts as "empirical adequacy" (e.g. must an empirically adequate theory predict the outcome of every experiment, or is a theory sufficiently empirically adequate if it instead predicts a range and probability distribution of outcomes?) can and will vary from theory to theory.

We are generally quite successful in interpreting ourselves as able to act according to our beliefs and with intent. To the degree that other systems resemble ourselves, we (attempt to) use a similar strategy. Just as we are successful in interpreting the outputs of our own system as intentional actions, we are successful in interpreting other humans in the same way. Generally speaking, we "read off" others' beliefs from their actions (of which utterances are a species) and by understanding their beliefs, can explain their past actions and predict future actions. Although we do not have direct access to others' thoughts or to the intentionality of their actions as we do in our own case, the ABI model is enormously successful in interpreting other humans. When Sahar says that dogs are friendly, I attribute to her a belief that dogs are friendly and expect that her future actions will conform to a rule that states as much. When she reaches her hand to a dog's ears and rubs them, I interpret the data (a woman approximately 5'6" tall rubbing a medium-sized canine about the head) as a minded agent possessing an intention to rub dog-ears, performing an action that will carry out that intention, and a set of beliefs ("Dogs are friendly," "It is good to pet dogs," etc.) that guide her actions in various dog-related scenarios. Interpreting her as a minded belief-system allows more thorough explanations of past and current behavior, and better prediction of future behavior, than competing systematic interpretations (i.e. interpretation as a causal system, deductive system, quantum system, geological system, planetary system, etc.).

Although the ABI model is quite successful in predicting and explaining human behavior, it has a fairly significant downside. Interpreting a system as a belief system requires ontological commitment to the agential character and mindedness of that system. In our own case, we have very good evidence for the existence of our minds, as noted by Descartes. In the case of other humans, we are quite comfortable in taking them to be minded because they are so easily

recognizable as similar to ourselves. In the case of animals, machines, and inanimate objects, the case is not so obvious. Although we are successful in describing the orbits of the planets in terms of gravity, mass, and so on, describing the system in terms of beliefs, desires, and actions does not yield success above and beyond purely mechanistic explanations. Natural processes aren't minded, even if they *are* law-like. Water flows downhill not because it *intends* to do so, but because gravity is sufficient to overcome the coefficient of static friction between the river and its bed. The flowing of the river might constitute an *event*, but it is surely not an action in the intentional sense of the term. Similarly, we can perfectly explain my coffee mug as possessing a mind and at all times thinking to itself "I am in the best possible spot to be." It then forms an intention to remain stationary, and indeed carries out this action. When I move it to a new location on my desk, the process repeats. But this interpretation of the stationary qualities of coffee mugs is simply *wrong*.

What, though, stands in the way of my attributing mindedness to the river and my mug? In Book V of the *Odyssey*, Odysseus prayed to the river – what was his mistake?[8] The mistake made when we anthropomorphize that which has no mind is multiple. In the first place, we become ontologically committed to entities – minds – that do not exist. It is, of course, generally better to only posit minds to those things that *really do* have minds. We know that rivers and coffee mugs do not have minds because we do not *need* minds to explain their behaviors – minds are superfluous. I can explain and predict the movements of my mug perfectly and robustly without any appeal to minds or intentions or beliefs, and I can do the same for rivers. I *cannot*

---

[8] In Odysseus' case, of course, he made no mistake, for the river god hears his plea and "soft receives him from the rolling sea" (Homer). If we assume that the world is not as Homer described it, though, and does not feature gods (river and otherwise), praying to rivers is not likely to be efficacious. Even if Odysseus' action was correct, we would be mistaken to mimic him.

explain other humans' actions without the apparatus of minds and beliefs and actions, however. Any sufficient explanation of human behavior requires mindedness and the corresponding ABI model.[9]

In the case of animals and computers, the line between minded and non-minded is rather blurry. Although the current state of AI is not sufficient to carry on a conversation as with a human being, if such a program were to exist and able to articulate its beliefs, desires, and so on the failing to attribute it mindedness seems rather odd. If the hypothetical program were run by simulating a human brain, failing to attribute it desires because its brain existed among NAND gates and magnetic platters instead of neurons seems rather silly, particularly when we take into account that alien mindedness – whose implementation might be wholly different from the architecture of the human brain – is not particularly controversial from a conceptual standpoint. The most salient criterion for attributing mindedness is whether a system's outputs are best predicted by such an attribution, not particular vagaries of the physical implementation of that mind.

What, then, of the *current* state of artificial intelligence? Chat programs can more-or-less simulate human conversation, and can at least some times fool humans into thinking that they are minded, that they have beliefs and desires and perform actions with intentions. What of animals? At present, humans are better at using language than any other species (or construct).

---

[9] This is a substantial, if unargued, thesis. I take it that at the end of inquiry the natural kind *mind* will be necessary to adequately describe human subjectivity and the first-person sensation of freedom to author and articulate one's own actions. If it turns out that human minds, sensations, and so on can indeed be fully explained and understood without reference to a particular mind or particularly mental properties, then there will be no need to speak of minds, mental properties, and so on and so minds and mental properties would no longer be atoms in a full description of all empirical data. If this is the case – if the eliminative materialist is correct - then they do not exist as abstract objects. Of course it may still be *useful* to talk about minds and mental properties, so long as we understand that they are shorthand for particular physical states and have no ontological import).

We attribute mindedness to other things, but it is not clear whether they can attribute mindedness to us. In this sense, humans are special. But the actions of a dog that systematically searches for a hidden toy, an orangutan that uses tools to find food and to shelter itself from rain, and a chimpanzee that takes pleasure in causing suffering do not seem to admit of a wholly mechanistic interpretation – there needs, it seems, to be *some type* of mindedness, belief, and intentionality in the mix even if it is not quantitatively (or perhaps even qualitatively) like our own.

To this end, individual interpreters are free to choose the extent to which they accept ontological commitments of using the ABI model to understand dogs, orangutans, and elephants in its literal meaning, and the extent to which they take its application to non-human systems to be analogies or a homonymous use of terms.

Although it is by no means certain that non-human systems do not have minds in the fullest sense of the word "minds," it is obvious that we can best explain and predict dogs' actions in terms of figuring out their desires (or "desires") and then figuring out what they think (or "think") will be the best way to satisfy those desires. Dog behaviors are governed by rules, and at least some dog behaviors seem at least partially intentional, to the extent that calling them "actions" does not *prima facie* seem to be a category mistake. But the dog-mind that we use to understand and predict dog behavior is quite different from the human-mind that we use to understand and predict our friends' behaviors. We expect quite different things from humans and from dogs, and one generally does not scold a golden retriever for failing to make deductively valid inferences. The ABI model that we use to understand the outputs of dog-systems - ABI_{dog} – is radically different, if structurally similar, to the model we use to explain and predict the outputs of human-systems. It seems clear that dogs have beliefs, but these beliefs

are not very much like ours (though dog-beliefs are certainly more similar to human-beliefs than the behavior-rules used by other less intelligent animals like cats and grasshoppers). Dog-action is better explained with appeal to dog-beliefs and dog-minds than without, but we do not expect or predict that dogs-minds and human-minds make the same decisions, care about the same things, and so on. A conservative position wherein dogs have beliefs, but they are only beliefs in a derivative and charitable sense seems to be the most reasonable. In cases such as tables, rivers, and hornets the ABI model adds little to nothing to our understanding of the relevant phenomena, and may safely be discarded (though this position is, of course, open to revision). In the cases of elephants and orangutans and New Caledonian crows, the jury is still out as to whether they have minds, "minds," or something else entirely.

Of course, philosophers are not solely worried about whether non-human animals, computers, and so on have beliefs. We also worry ourselves about whether groups of humans can be said to have beliefs (see e.g. Gilbert, 1987; Gilbert, 1994; Wray, 2001; Gilbert 2004; Bouvier, 2007). The conceptual model of belief that I am using applies just as easily to groups as it does to dogs, cats, and thermostats, which is to say that it is not obvious how well it applies at all. To the extent that we can look at a group and plausibly tell a story wherein that group takes actions as a group, and to the extent that these actions seem to be governed by rules, then it looks as if groups have beliefs. It is a problem, though, that groups fairly clearly do not have a single mind, at least not in the same way that I have a mind and you have a mind. But of course, dogs do not have minds like I have a mind and you have a mind, either. They have *something*, though, and there is a degree to which it is appropriate to be charitable in mind-attribution.

It is a fact that we never get direct access to others' mental states. Our experience of other minds is solely through abduction: the best explanation as to why this collection of

biological components that I call my wife moves and thinks and acts in a way that is looks to be *like me* is that she *is indeed like me*. She has a mind, and thoughts, and hopes, and beliefs. When it comes to groups, they sure *seem* to take actions. To be sure, groups can make announcements and condemnations; can purchase, sell, and own property; and can come into and go out of existence. Indeed, some groups are recognized as legal persons in the United States. If a group can sue or be sued, it certainly looks like it acts. If it acts, it acts for a reason, or with intention. The method by which it acts is governed by certain rules that describe expected outcomes of actions and so on, and we call these rules "beliefs." The "mind" that is posited in treating groups as ABI-systems is just what we need it to be, and no more: that of which we predicate the group's beliefs. I believe that human consciousness is fundamentally linguistic, but I have my doubts that dog-minds are so regimented. Group "minds" may or may not be linguistic, and I make no empirical claims as to their nature. I simply note that, upon recognizing belief-attribution as interpretation of a system according to the ABI-model, there is no impediment to so interpreting groups of humans (or, for that matter, ants or fish or wolves).

Of course, the *relation* between individual and group is a separate question entirely. Gilbert (1987) argues quite forcefully that the member/group relation is *non-summative*, in the sense that a group's belief is not the sum of beliefs of all its members. Thus a group could take an action that none of its members think is the correct course of action, but each member takes to be the second-best option. This is a welcome result, as if we admit that there is something to a group's *being a group* over and above its members, this makes room for the type of mindedness attributed by the ABI-model.

I wish to remain agnostic as to whether the ABI model properly applies in any particular case, as it seems obvious that its applicability is not a binary matter. Rather, the suitability of an

interpretation to a particular matter is scalar. Dennett (1995) takes a very wide view for what counts as a belief, and counts thermostats and amoebas as sufficiently minded for belief-attribution. Davidson (1984) and others think that thought, belief, and so on require language (though both are generally amenable to the claim that belief involves taking a particular stance towards a system). While I am sympathetic to the claim that *human-like* thought requires something that we can recognize as language, this does not say much. Dogs have no language, but they seem to approximate thought or something like it, and they seem to have something close to belief. Very little seems to depend upon the question as to whether that belief-like thing that dogs do really is *belief* after all and not schmelief or doggy-belief or what-have-you.

While I am largely in agreement with Dennett (1983, 1985, 1987), I suggest that in taking his intentional stance towards – or, more precisely, applying the ABI model to – a system, we are ontologically committed to at least some type of mindedness for that system, the mind in which particular beliefs inhere. Thus it seems flatly *wrong* to me that his amoebas and thermostats (1995) are good candidates for application of the ABI-model. But this is of little import: speaking of amoebas and thermostats as having beliefs is being quite liberal with what counts as "minded enough" to warrant interpretation as an ABI model, but my disagreement with Dennett is largely semantic and not likely to bear philosophical fruit.

Though I am reluctant to attribute belief as freely as Dennett in his (1995), I do not wish to go too far in the other direction: where Davidson (1982, 1984) thinks that language is a necessary condition for having beliefs and that belief-talk only makes sense in the context of triangulation with others and an outside world, I only say that we need some system to interpret according to the ABI model. That system could be ourselves, and while of course we need

language to apply the ABI model (since it is articulated linguistically), there is no reason that we could not so interpret non-linguistic systems.

I agree with both Dennett and Davidson that belief is most importantly an attribution to a system that allows us to better understand and predict that system. In Peircian terminology, belief is a *rule for action* of that system, where *action* is simply understood as the intentional output of a minded system, which is to say a system interpreted according to the ABI-model.

3. Features of Belief as a Rule for Action

To say that beliefs are rules of action does not imply that they are non-propositional. To believe a proposition *p* is indeed to take the attitude towards *p* that *p* is the case, and *doubting p* is being unsure as to whether *p* is the case. A useful start to a formalization of this characterization would be to say that an agent believes *p* if she holds the attitude towards *p* that it is the case with a credence of greater than .5, and that she doubts it if her credence is .50 or less. Any individual (or even humanity in general) might take an arbitrary credence of greater than .50 as sufficient for transforming belief to doubt, though it is a conceptual impossibility that one might have credence of less than .50 with respect to a proposition and still expect that it is the case in any meaningful way. The concept of "taking something to be the case" demands that any taking-to-be-the-case attitude towards a proposition with credence of .50 or less counts as a doubt (and not a belief), but the floor for counting-as-belief can (and perhaps should) be set higher.

Doubts, on the other hand, are acknowledged absences of rules for action.[10] In a state of doubt, an agent is unprepared for action precisely because she has no rule to guide her behavior. She is hesitant to perform actions related to her doubt, because she does not expect that the results of her action would conform to her desires. Depending on the severity of her doubt, it may be resolved in a matter of seconds – as when she glances into her purse to ensure that her wallet is present – or might take weeks or years to resolve, as when after long toil she proves the impossibility of contingent identity.

Necessary to the above description of belief is a weak notion of the psychological transparency of belief. This (controversial) thesis states that it if I believe *p*, it is psychologically accessible to me that I believe *p*. Although this maxim sounds strange at first, it is quite plausible in light of belief's role as a rule for action. The most obvious way for a belief to be psychologically accessible is for it to be *used*, which is to say that it is *accessed* (though certainly not on a conscious or reflective level) and implemented in action. If a belief is *inaccessible*, then it must be the case that it is impossible for an agent to *notice* that they have a belief. This sort of belief, as a rule of action, is most saliently identified by the actions of an agent[11] – insofar as it is possible for an agent to act on her belief and for a particular proposition to be posited as the rule to which her action conforms, that belief is psychologically accessible in the weak sense needed.

This is not to say that an agent cannot be *wrong* about her beliefs. To believe *p* is to have a general rule for action such that one will act as if *p* is the case. If, as a general rule for action, I behave as if dogs are extremely dangerous – if I enter an intense state of fear, seek to leave the animal's immediate presence, and so on – it is appropriate to say that I *believe* dogs are

---

[10] That is, if Sahar is consciously unsure as to whether *p* is the case, then she doubts *p*. If she has never considered *p* one way or the other, then she has no attitude whatsoever towards *p*.
[11] That is, by application of the ABI model to the agent's behaviors

dangerous even if I sincerely deny the proposition when it arises. In such an instance, I *believe* that dogs are dangerous, but I also *believe that I believe that they are not dangerous*. My belief that dogs are dangerous is psychologically accessible to me in that it is able to form the basis of my actions. If I understand what belief is – that it is a rule for action – and I therefore understand that the best guide to an agent's beliefs is his or her actions, then I can diagnose myself as believing that dogs are dangerous if I am in a particularly reflective mood. That I have an incorrect understanding of belief or a deep-seated commitment to my own machismo or what-have-you that blocks this recognition does not undermine that the belief is accessible in principle.

Consider the following scenario: After much thinking and mulling over, I reveal that a moral proposition *m* is consistent with my background moral principles, facts about the world, and so on. It would not be absurd to say that I believed *m* all along, and simply uncovered that belief by attending to the various implications of my background beliefs. While such a manner of speaking is certainly natural it is not strictly-speaking correct in the absence of a proper belief (that is rule for action) that *m*. The process of "mulling over" is better characterized not as uncovering a belief that had always been present, but rather as *inquiring* into whether or not *m*. That is, for some reason or other, I am led to doubt whether *m* is really the case.[12] The most obvious way for this to occur is for me to realize that, in *m*-relevant situations, I am unsure as to the proper course of action. That is, I notice a lacuna in my beliefs with respect to *m*-situations, thus developing a doubt with respect to *m*. This doubt leaves me unprepared to act with regards to situations relative to *m* – if an *m*-relevant situation were to arise, I would not know how to act

[12] For this to make sense, I must have never encountered or acted in an *m*-relevant situation. If I had, then whatever rule I acted by would be the relevant belief about *m*.S

(and it is my realization that I have no particular habit of action with respect to *m*-situations that births the doubt). This state, like all doubt, is uncomfortable. It is worrisome that an *m*-situation might arise and I would be unprepared for it, and so I seek to decide whether or not I should believe that *m* – inquiry begins with a doubt. The process of examining the role of *m* within a network of physical, moral, and methodological commitments constitutes inquiry in this case; when I at last decide that *m* is consistent with my background commitments, and ~*m* is incompatible with them, then I have settled the question as to whether or not *m*; I judge that *m*. This judgment – the result of inquiry – is a representation that *m*, an initial manifestation of my settled belief that *m*. While the judgment lasts only as long as it is occurrent in my consciousness, the belief remains in a latent form as a rule for action, or habit. After I have settled my belief that *m* (which in most instances will involve a judgment that *m*), I will be disposed to act according to newly-inculcated belief that *m* in *m*-relevant situations.

In situations dissimilar to the above, where I *am* prepared for action with respect to *m*-situations, then I simply believed *m* but did not believe that I believed it. Of course, I *also* did not believe that I believed ~*m*; I simply lacked any meta-belief with respect to *m* at all.

4. All belief is dispositional

The lack of complete psychological transparency of belief is best understood through the distinction between belief proper (which is a rule for action) and judgment or occurrent thought. This distinction is closely related to the contemporary distinction in epistemology between occurrent and dispositional beliefs. Just as occurrent belief is a manifestation of dispositional belief, judgment is both a manifestation of and the reinforcing of the "cerebral habit of the

highest kind" that is belief proper (Peirce 1880, p. 16). As a dispositional belief is one that we *would* occurrently believe if it were to come up, when we believe *p* in my sense, we are in a state such that *ceteris paribus* we would judge that *p* if we had occasion to do so.[13] A judgment – or thought – is the conscious process of stating *p* to oneself.[14] It is an action, and it is also the most internally obvious manifestation of belief. It is particularly useful that thought – or judgment – is propositional. Given the close connection between *having* a rule and the most obvious psychological implementation of that rule (which itself gives the rule much if not all of its propositional structure and allows us to talk about it), it is not surprising that we often talk about judgments in terms of their being occurrent *beliefs*. Although this category mistake is quite easy, the distinction between beliefs (which are rules posited by an observer of a system to understand, explain, and predict its outputs, noting that the observer might be the system itself) and judgments (which are outputs of that system, though they are only directly accessible to the system itself via the mind) is crucial. While seminal papers in the metaphysics of belief maintained the status of occurrent belief as belief (see e.g. Audi, 1994 and Lycan, 1986), more recent authors (e.g. Schwitzgebel, 2010) tend to use my preferred terminology in strongly distinguishing (occurrent) judgment from (dispositional) belief. Although I believe my usage of the terms "belief" and "judgment" to be the clearest way to distinguish the two phenomena, as long as my interlocutor agrees that there *really are* two phenomena, we need not argue over the semantics of whether a judgment is "really" a species of belief.

---

[13] This assumes that we do not also believe ~*p*, in which case we might judge ~*p* while behaving as if *p*. Hence the *ceteris paribus* clause.

[14] It is not implausible that groups, animals, etc. that we interpret according to the ABI-model do not make judgments. This is unproblematic, since a judgment is just one of many types of action that an ABI system can perform. It is also not implausible, of course, that groups, animals, and so on do have something that plays a similar role as judgment, even if it is e.g. non-linguistic.

Despite the similarity of their surface structure, one might assert a distinction between dispositional and pragmatist belief on the following grounds. Since belief, for the pragmatist, is the end of a process of inquiry occasioned by doubt, it might seem that certain habits we have cannot count as belief, since they have never been doubted. They are, so to speak, baked in to our cognitive and cultural practices and are adopted without question, without reference to a moment of doubting.

I deny the above. Doubt is that state wherein we are unsure what to do; it need not be soothed by weighing all available evidence. While the scientific (that is, fallible and evidence-responsive) method of inquiry is the *best* method because it is truth-conducive over repeated applications, doubt can be driven away simply by taking something to be the case. To be sure, if we are informed that our belief formed in such a manner was in fact merely assumed without evidence, this will usually occasion us to doubt again, but inquiry as such need not be a robust cognitive process that weighs evidence for and against a claim. The method of inquiry that does in fact weigh all available evidence is preferable in that it is truth-conducive and our belief-habits so formed are unlikely to be the basis of failed action, but the process of inquiry is the transition from a hesitancy to the formation of a rule for action. Thus the child that blindly accepts the values and mores of his society or the adult who is a victim of confirmation bias cannot be accused of failing to believe, but rather of failing to inquire in an appropriately truth-conducive manner.

If I am correct, and we can rightly analyze pragmatist beliefs as dispositional beliefs, an immediate problem arises. Peirce counts as the first of belief's three essential properties that it is "something that we are aware of" (Peirce 1878, p. 29). But, as we have seen, a Peircian belief is

dispositional, it is a rule for action or a habit. Surely, though, we are not aware of the totality of our physical and intellectual habits.

We can make room for the above objection. Those habits we have of which we *cannot* be aware are not beliefs. A belief is not *just* a cognitive habit, but a cognitive habit that is by definition accessible in an extremely minimal way. Just as the real is partially defined as that of which we can know, belief is defined as that of which we could become aware. This is not to say that a man is aware of his belief that *p* at all times in which he believes *p*, but rather that if a man believes *p*, he can access his belief that *p* in the sense that he can use it as a basis for judgment, action, expectation, and so on and can (conceptually speaking) posit a belief that *p* from his past actions that assume *p* is the case. If it is impossible in principle for *p* to be the rule by which *any* action is undertaken (whether that action is a physical action, a verbalization of a proposition, or a thought), then *p* is not a belief. Put another way, for a belief to be a rule of action, there must be the possibility of some action that is an instantiation of that rule. Without the possibility of instantiation (e.g. through judgment, expectation, etc.), there is no sense in which the proposition is a rule, and therefore no sense in which it is a belief.

Pragmatist beliefs coincide with the dispositional beliefs of the contemporary literature, and pragmatist judgments with occurrent beliefs. A belief is a rule for action. Beliefs are cognitively accessible in that they can be (but need not always be) called into question in light of contradictory evidence, frustrated action, and so on, and that they can be (but need not always be) acted upon. When a belief is directly cognitively accessed and affirmed or denied, that affirmation is a judgment that *p* (or not-*p*). While beliefs are latent in a sense, they are manifest in judgment and action; we can say that the belief that is manifest in a judgment that *p* represents *p* as the case, and it is by virtue of representing *p* as the case that we take *p* to be an appropriate

grounds for action (that is, we feel comfortable in acting on the basis of *p*). The propositional nature of a belief is manifest both when it is judged to be the case (that is, when the relevant proposition is directly cognitively accessed) and when it is used as a basis for action. In the case of action, it is of course unclear what *particular* belief is at play – we could plausibly say that avoiding a bonfire involves various beliefs about fire being very hot, extreme heat being painful, pain being something to avoid, and so on. As such, it seems prudent to consider judgments – explicit cognitive occurrences of "Fire is very hot" and "Extreme heat is painful" and so on – as the paradigmatic cases of propositional belief. Nevertheless, beliefs themselves are propositional in the sense that they involve a (perhaps latent) representation of a proposition whose propositional nature is revealed only vaguely in action, and explicitly in judgment.

5. A puzzle about contradictory beliefs

The lack of complete psychological transparency of beliefs leads to some rather interesting puzzles. Consider that when presented with a dog, Ralph tends to sweat, act nervously, and often comes up with excuses to leave the room. Consider also, though, that if we ask Ralph what he thinks of dogs, he sincerely and honestly says that he is not afraid of them, and that dogs are indeed good friends. On the one hand, Ralph acts in a manner consistent with the proposition "(All) Dogs are dangerous," but on the other hand his actions (including his thoughts and utterances) are consistent with "(All) Dogs are not dangerous." Although these propositions contradict one another, Ralph acts according to each. Ralph believes that he believes that dogs are not dangerous, and so in reporting as much he performs actions consistent with such a position (primarily his thoughts and speech acts). But his other actions (leaving the

room, positioning other individuals between himself and the dog, and so on) reveal that Ralph does indeed take dogs to be dangerous. Is Ralph guilty of holding contradictory beliefs, or should we posit some other belief for him, something like "Dogs are non-dangerous with regards to speech acts and thoughts, but are dangerous with regards to physical interaction."

The two possibilities are thus that Ralph believes both *p* and ~*p*, or that his belief is a much more nuanced rule with a rider to the effect of "Affirm belief in the non-dangerous nature of dogs even when acting against according to the opposite." This is precisely the debate between Socrates, who thought all moral failing to be a sign of ignorance as to what is good, and Aristotle, who thought that the state of incontinence was defined by having the proper beliefs about what is good but being unable to conform one's actions to that ideal. What Aristotle would call an incontinent belief – belief, say, that cigarettes are bad for me while smoking – Socrates would characterize as my not *really* believing that cigarettes are bad for me, though he would surely admit that I believe that I believe that cigarettes are bad for me.

The Socratic position denies Ralph's belief in the relevant proposition when his actions fail to conform to it (for a contemporary version of this view, see Hunter, 2011). To the contemporary (or, for that matter, Aristotelian) ear, this seems too strong – surely belief is a weaker notion that does not demand complete consistency in action. The position I call "Aristotelian" says that Ralph believes that dogs are friendly, but that he does not act on it (Zimmeran, 2007; Gendler, 2008a; Gendler, 2008b). Let us admit, ceding a small point to Socrates, that there is something special (indeed *ideal*) about behaving in complete accordance with one's beliefs. It is best for us, as rational inquirers, that we have consistent beliefs and that we act in accordance with them. But let us cede a point to Aristotle as well: incontinent beliefs

really *are* beliefs, in that they provide a rule for action in some minimal capacity. That minimal capacity is, at least, sincerely affirming the proposition in question.

Like all things, beliefs are theoretical entities. They are individuated by their utility in explaining phenomena. It might not be clear whether I believe "Fire is painful to touch" or "Fire is ritualistically impure" from a single instance of my avoiding a flame. It is not always even clear to *me* whether this is the case. A belief is a rule for action, but I need not be aware of it, and so I need not believe that I believe it. I can even believe that I believe its opposite. It is quite difficult to tell what belief corresponds to what action – the totality of one's beliefs only come about from repeated observation and prediction (hence the infamous "I'm not racist but…" preface to so many racist comments). In Ralph's case, who says "I do not believe that dogs are dangerous" while avoiding dogs and displaying physiological and psychological fear responses, it is obvious that he is mistaken: he certainly believes that dogs are dangerous. What is less obvious is whether he *also* believes that dogs are non-dangerous (and thus has contradictory beliefs *a la* Gertler, 2011), or whether he has a false belief about his dog-beliefs and that is the end of the story. Is Ralph's minimal belief (as evidenced by his speech acts) that dogs are not dangerous enough to generate a contradiction?

To accuse Ralph of holding inconsistent beliefs, it must be the case that an inconsistent set of rules is the best explanation of his actions. However, for any apparently inconsistent belief-set that explains a set of actions, we can generate some other belief-set that is consistent and explains just the same actions. This second belief-set will, in all instances, explain *better* than the first, for its rules are never violated. To re-visit the example of dogs and the phobia thereof, Ralph acts sometimes as if the guiding principle of his actions is "Dogs are dangerous," and acts sometimes as if the guiding principle of his actions is "Dogs are not dangerous." Both

sentences, though, are occasionally wrong (thus the necessity of positing the contradictory position). However, the alternate rule "Dogs are dangerous at times t_1, t_2, and t_3 and are not dangerous at times t_4, t_5, and t_6" is always satisfied, and fully describes Ralph's actions. It thus seems that it is never legitimate to posit that an agent holds contradictory beliefs, because we could always give a single rule that, in the most deviant case, is a long conjunction of every action he performs. What, then, could privilege the positing of contradictory beliefs over the positing of more specific but consistent beliefs?

To solve this puzzle, we must remember that any particular belief is an explanans for some particular (possibly singleton) set of actions. In attributing a belief to someone, we attempt to make their actions intelligible to us while preserving their character as an agent that is essentially *like us*. In this attribution, there is an element of charity: we do not *ceteris paribus* suppose that Sahar took her umbrella to work with her because she believes that umbrellas ward off marauding dragons. We similarly do not suppose that she took her umbrella with her to work because it was May 9 or because her dog is a German shepherd. The lookup-table method of belief-attribution mentioned above, though consistent, attributes beliefs to an agent that are essentially *irrelevant* to the situation at hand.

The irrelevance of dog breeds and the date to bringing an umbrella to work becomes most obvious when we recall that while we can *arrive* at beliefs in myriad of ways, beliefs are *tested* by the success or failure of the actions that conform to them. It is quite hard to imagine precisely *how* Sahar could come up with "Umbrellas are necessary on May 9" or "Umbrellas are necessary when your dog is a German shepherd" without these *actually* being at least somewhat successful rules for action with respect to bringing umbrellas to work. Since we know that Sahar's dog will *always* be a German shepherd (but one doesn't always need an umbrella) and that one doesn't

need an umbrella on the ninth of every May, attributing such beliefs to her is untenable. In attributing beliefs to others, we also attribute to them histories of success and failure in action that reflect our own to a greater or lesser degree. Lookup-table methods of belief attribution fail to (even implicitly) have any plausible way that an agent could arrive at such a belief system.[15] In a phrase, they are genealogically untenable. Thus while it is not impossible that someone *really does* use such a system, actually *attributing* the use of such a system to them is always a bad idea.

Although this response can help us in forming some constraints with respect to belief-attribution, it does not fully resolve the problem of attributing contradictory beliefs. While it might be untenable to attribute overly-complex or conditional beliefs to an agent without good evidence for doing so instead of attributing simpler, more general beliefs to them, the problem of contradiction-attribution is that the more complex consistent belief *better* predicts an agent's actions than do two contradictory conjuncts. The former has always been followed, while whenever one conjunct of the latter is followed the other is necessarily violated. How, then, are we to make sense of Ralph's position?

We can resolve the puzzle by making note of the fact that belief networks are extended in time. When Ralph has inconsistent beliefs, he acts by a rule that p at t_1, and act by a rule that $\sim p$ at t_2. It is not implausible that the belief that p is simply ineffective in guiding action at t_2. Indeed, if the belief that p can be said to be held at all, it *must* be inactive or ineffective or otherwise divorced from Ralph's action. After all, if the belief that p were active or effective or

[15] Moreover, a lookup-table sort of belief system fails to have *rules* for action in any meaningful sense: each action is time-indexed, and thus each "rule" is instantiated only once. I fail to see how a rule can be a rule in any meaningful sense if it can only be instantiated a single time. Indeed, single-use rules negate the very *point* of positing rules for action (i.e. beliefs) to agents: predicting their future behavior.

action-guiding, Ralph would act by *p* (and not by ~*p*) at t_2. Given that the agent does indeed act that ~p at t_2, something or other must have happened between t_1 and t_2. We call this breakpoint – from a belief network that guides action towards *p* to a belief network that guides actions towards ~*p* – changing one's mind.

Let us examine Ralph before and after his actions. Let *p* be the rule that (all) dogs are dangerous, and ~*p* be the rule that they are (all) good to pet. Let t_1 be today and t_2 be tomorrow. Let us also assume that Ralph has never interacted with dogs (save what's necessary to form the relevant rule for action) before today, and will never again interact with dogs after tomorrow. He has but two discrete instances of doggie-interaction to evaluate.

Today, Ralph behaves fearfully around dogs. By hypothesis, his rule for action is that dogs are dangerous, and the most reasonable evaluation of his actions is that Ralph *believes* that dogs are dangerous. A day passes, and sometime during the night Ralph's dogs-are-dangerous rule for action dissolves and a new rule is formed: "Dogs are good to pet." Ralph wakes up, finds a dog, and pets it. Given that this is his only dog-relevant action for the rest of his life, we can plausibly say that yesterday (and before) Ralph believed that dogs were dangerous, but that today (and in the future) Ralph believes that dogs are good to pet. We say that between yesterday and today, Ralph changed his mind – the breakpoint marks a significant shift in observable action, and thus an observer's best bet (whether that observer is Ralph himself or a second party) is to posit certain rules to Ralph's behavior in order to predict his future actions.

There is an alternate interpretation of Ralph's actions that obviates the need for positing a change in mind. An observer could say that all along (and in the future), Ralph has believed "Dogs are dangerous until and including t_1, and dogs are good to pet from t_2 on." Both ways of explaining Ralph's behavior do so equally well, but the second (singleton) belief set seems to

have the advantage that it explains Ralph's behavior with a single rule, even though the first belief-set is intuitively more plausible.

The superiority of the first belief-set lies in the facts that we don't think that the nature of dogs changes according to time. That is, we don't think that t_2 was some watershed moment that forced all dogs to be good-to-pet instead of dangerous. That would be a silly belief, and so hypothesizing that Ralph believes it and acts by it is uncharitable.[16] This is a mark against the second belief-set. The second mark comes when we simply *ask* Ralph what he believes. "Ralph," we say, "were dogs dangerous a month ago?" If Ralph is a sensible sort, he will say "No, but I believed that they were." The second belief-set must be thrown out because all-told, it is a rather bad explanation of Ralph's behavior.

Between t_1 and t_2, Ralph changes his mind. Before and after the breakpoint, Ralph has consistent belief sets – we can only attribute to him inconsistent beliefs if we insist on taking Ralph's belief set to be the conjunction of all his beliefs both before and after he changes his mind. At no instant, then, does Ralph have inconsistent beliefs, *contra* Gertler. Rather, his beliefs are inconsistent over time, but we can make sense of these inconsistencies by positing changes in mind – that is, in locating points in time before and after which Ralph's actions (and this the rules that guide them) differ dramatically.

Ralph's case is, of course, quite simple. We can make it more complex by adding more changes of mind from the p rule to the $\sim p$ rule and back again. Given enough changes of mind, it might be impossible to tell if Ralph will pet a dog or run screaming. Worse, he might run away while screaming "Dogs are good to pet." Would these be cases of inconsistent belief?

---

[16] Indeed, this is just a small-scale application of the lookup-table method of belief-attribution discussed above.

I think not. In the first case – where Ralph is simply unpredictable – his actions (and the corresponding beliefs) are unpredictable, and seem to constantly be in flux. At any instant, however, Ralph presumably has *some* rule or another by which he would act with respect to dogs. It is, of course, possible that Ralph has *no* beliefs regarding dogs and whether or not they are dangerous, and that a rule is formed only upon the occasion of seeing (or being asked about, or thinking about, etc.) a dog. In either case, though, Ralph has a consistent belief-set at each instant, but exhibits significant instability in his beliefs over time.

In the second scenario, in which Ralph *acts* as if dogs are dangerous while *saying* (which is itself acting) that they are not, we can make sense of his behavior by positing a meta-belief: *that* he believes dogs are good to pet. Thus Ralph believes that dogs are dangerous, and believes that he believes that dogs are good to pet. While this leads to behavior that is perhaps difficult to predict and rather puzzling, Ralph's problem is not one of *inconsistent* beliefs, but rather of a false belief *about* his beliefs.[17] He believes that dogs are dangerous, and he believes that he does not believe this. In the second case, he is of course wrong, but error is a far cry from inconsistency.

Inconsistency of beliefs, then, is best analyzed as inconsistency of beliefs over time. It is best analyzed not as an agent simultaneously holding two contradictory rules of action, but rather of the presence of these contradictory rules separated by a temporal gap. Indeed, genuine inconsistent belief at an instant would require an inconsistent set being the very best model for an individual's beliefs at that time. Given that a belief-set is always open to addition (including the addition of incorrect beliefs about beliefs to make sense of statements that contradict one's

---

[17] And so I ultimately side with Aristotle.

actions), we can always posit a consistent belief-set to make sense of an agent's (including our own) actions. Thus the puzzle about inconsistent beliefs is resolved not through dissolution, but rather by admitting that when we speak of inconsistent beliefs, we speak not of an agent's belief set at an instant, but rather extended through time, and that an agent that would commonly be said to hold inconsistent beliefs is better described as exhibiting an (perhaps pervasive) instability of beliefs.[18] Thus I follow Rowbottom (2007) in modelling Ralph as wavering between his belief that dogs are friendly and his rejection of that proposition.

It is worth noting that there is an alternate route: Schwitzgebel (2010) models similar cases as "in-between" or "vague" believing. I am generally comfortable with this strategy, as it follows naturally to model degree-of-belief in uncertainties along such lines.[19] For example, we might say that Ralph believes "Dogs are friendly" with credence .55, "Dogs are dangerous" with credence .45, and that the best overall description of him is as an "in-between" believer of either of those propositions. But I do not take such a position to contradict my own: on a granular, synchronic level, either Ralph's recent behavior is best explained by (i.e. conforms best with) "Dogs are dangerous" or "Dogs are friendly." It is true that diachronically – through many apparently-contradictory actions – Ralph is not best described as determinately believing either proposition. Schwitzgebel notes that across time, either definite proposition does not fit well. I maintain that he is correct, and that the *reason* that neither proposition is a good rule for Ralph's

---

[18] There is, of course, an easy and familiar way to model such instability: degrees of belief, often described in terms of what odds one would be willing to accept as a fair (i.e. zero expected-value) bet. I take it that we really *do* believe propositions to various degrees, and will indeed make use of degree of belief in Chapter 4.

[19] Schwitzgebel is clear that such modeling is not sufficient, but this seems to be a failing of the Komogorov calculus used in Bayesian explications of degree of belief. Possibility calculi, generally formulated specifically calculated to account for vagueness and uncertainty, would do nicely. See, for example, Dubois, Lange, & Prade (1991). The fact that one specific logic is unsuited to the articulation of degree of belief is no reason to throw out the concept all together. A much more rewarding strategy would be to simply use a better logic.

action is that his actions fluctuate between acting according to "Dogs are friendly" on the one hand and "Dogs are dangerous" on the other.

Rather than use an instability-analysis of contradictory beliefs, one might be tempted to posit the contradictory belief as eternally present, but to distinguish it in some fundamental way from other beliefs in such a way that one is not necessarily incorrect in holding such beliefs. This is the approach taken by Gendler in her distinction between beliefs and aliefs (2008a, 2008b, 2010). For Gendler, beliefs are obvious from a first-person view, while aliefs occupy that muddy ground between belief and instinct/reaction/habit. Aliefs are those things that we share with animals, that are affective, and so on and beliefs are something else. While I readily accept that some action-rules are obvious to us (even when we act by contradictory action-rules), and that some are not, Gendler's case for a conceptual distinction between alief and belief is not necessary if one adopts my understanding of belief. She says that beliefs, unlike aliefs, attempt to track truth. And it is true, of course, that beliefs (try to) track truth, insofar as a frustrated action leads (in the best case) to doubt, inquiry, and belief revision. Do aliefs not *also* track truth? The obvious and well-documented phenomenon of habituation provides us ample evidence to conclude that a repeated belief (or even a repeated acceptance) can "turn into" an alief. One favorite example is that of the Grand Canyon's Sky Walk (Gendler 2010, p. 261-262), in which people who are directly aware that they will not fall (thanks to the thick glass floor of the observation unit), but are still convinced that they will indeed fall. According to Gendler, the individual on the Sky Walk alieves that they will fall while believing that they will not. I grant Gendler that the scared Sky Walk patron alieves that they will fall. Why, though, does the maintenance worker not so alieve? She has cleaned the glass every day, and she has sufficiently habituated herself to *not* alieve that she will fall, because she has ventured on the glass enough

times to firmly entrench her belief that the glass will hold. Certainly aliefs are *harder* to dislodge than beliefs, but the position that they do not track truth implies that aliefs are not revised in the face of contradictory evidence, which is markedly contrary to fact. The first time I rappelled, I alieved I would die and was afraid, though I believed I would live. Through repetition, I have come to alieve (as well as believe) that I will live, and am not afraid to rappel. The most obvious conclusion is that aliefs *just are* beliefs, but they are particularly *deep-seated*, or difficult-to-access, or tenacious beliefs. They are beliefs that are easier to read off of our own actions than from our thoughts, but they, like beliefs, are just rules for action. Insofar as action is more successful when it conforms to the state of the world (that is when the action's rule is *true*) alief, like belief, attempts to track truth.

Gendler characterizes aliefs as distinct from beliefs in that aliefs are not responsive to the norm of accuracy (2010, p. 271). But this just seems wrong. She is certainly correct that some rules for action are deeper-seated than others, and that some rules for action may be more easily jettisoned in the face of evidence that contradicts them. However, the success of exposure treatment for phobias and the success of cognitive-behavioral therapy in general tell us that even our most foundational, difficult-to-access rules for action are in principle amenable to revision. Most importantly, while any specific alief might be practically outside the realm of habituation therapy and eventual revision, there does not seem to be anything conceptually problematic with holding aliefs to be revisable in the light of repeated contrary evidence. While I agree with Gendler that some rules for action are more difficult to change than others, once we properly understand beliefs as rules for action and judgments as a type of action, we see that the conceptual distinction between beliefs and aliefs collapses, and that aliefs cannot be a storing-house for what would otherwise be contradictory beliefs.

Beliefs, be they deep-seated and practically unrevisable or obvious and easily altered, are rules for action. Their most obvious psychological manifestation is their propositional occurrence in consciousness of thought (or judgment). Of course, one must believe before one can know. Indeed, the (plausible, I think) common view of knowledge is that one knows something when one believes it, when that belief is true, when has the appropriate justification for their belief, and when one is not overly "lucky" in believing as they so do.

6. *True* beliefs

Beliefs are rules for action most readily individuated by their most immediate propositional representation, by the thought that constitutes both the affirmation and a following of that belief. In believing *p*, we accept that *p* is an appropriate basis for action, and feel comfortable in using it as such. Although it is tautologous to say so, we take our beliefs to be individually true, though presumably we all expect that some of our beliefs are false. But what does a true belief look like?

In the previous chapter, truth was taken to be the end of inquiry as such – the *truth* is the set of propositions that form the appropriate basis for action in the limit of inquiry and leave no room for future (rational) occasion of doubt. As a system of propositions, the set of true sentences – of true judgment – carries with it a particular model. By "model" I intend that which is captured by "conceptual scheme" (in contemporary philosophical terminology): a language or vocabulary (set of variables and predicates with a grammar, etc.) L, domain D, and an interpretation function f which assigns to each individual variable in L an object in D, and assigns to every n-ary predicate of L a set of ordered n-tuples of elements of D. This is just to

say that if we assume that the proposition "All golden retrievers are dogs" is true, the interpretation function ensures that by "golden retrievers" we mean golden retrievers (and not hats), and by "dogs" we mean dogs (and not electrons). By virtue of the definition of "end of inquiry," we know that the whole of reality is accounted for in our set of sentences: all the objects in *D* have names in *L*. We also know that because the set of true sentences describes *reality* (and not a fiction), all variables in our model name objects in *D*.[20,21]

Let us call the model described above *M*. A true belief is rule for action whose corresponding propositional representation (that is, whose corresponding judgment) is a well-formed formula in *L* satisfied by *M*. It is not difficult to sketch what a true judgment looks like when all the data is in, at the end of inquiry: a true judgment is an occurrent representation of a proposition that an agent takes to be the case, and that proposition is in fact the case. The proposition so judged is true just in case it is satisfied by *M*. Because the proposition actually *does* represent the real, it is a suitable basis for action; when its representation has become a rule for action or a cognitive habit, it is a true *belief*.

---

[20] This is not to say that *L* contains no terms that do not denote objects in *D*, but rather that such terms do not show up in our set of true sentences (this is, after all, what it is for a sentence to be *true*)

[21] It is worth noting that there could be many objects that account for the same phenomena at different levels of analysis. That is, even though dogs and humans and so on might in a fully atomic language be best described as bundles of protons, electrons, etc. that are mostly made of empty space, it is still useful to talk of dogs as objects themselves. Thus a purely-physical account of a dog might only make reference to subatomic particles, but bundles of those particles could usefully be called "atoms," bundles of those atoms "molecules," and so on all the way up to "dogs," "packs," "ecosystems" and so on. So long as there is nothing about "ecosystems" or "dogs" stating that they themselves are atoms of the language, the model at the end of inquiry should be able to sensibly talk about artifacts and animals – and allow their full-stop existence as part of reality – so long as the use of the more macroscopic terms is theoretically useful. Since we have no idea just what the theoretical virtues would be at the end of inquiry (with the possible exception of empirical adequacy and parsimony discussed above), there is no pressing need to deny that intelligibility to human beings (and the corresponding inclusion of "dog" as well as "collection of subatomic particles" in our theoretical vocabulary) can itself be a theoretical virtue (we are, after all, the theory-makers). If one wishes to take a hardline stance and be an eliminative materialist, that position is of course open, but that is a commitment external to the theory that would shape its development, not a necessity of my understanding of truth and the end of inquiry.

And here a problem arises. The model of sentences that *really does* describe the real world is the set of true sentences. This is plausible enough, I think. We are, however, in the unfortunate situation of *not* being at the end of inquiry. *Our* beliefs surely do not map to the real world in the same way that an agent's beliefs at the end of inquiry so would. Even if we grant the implausible assumption that the *language* (that is, syntax and vocabulary) is similar between the model of my beliefs and *M*, surely the interpretation function (what sorts of things in the real world correspond to our predicates, terms, and variables) is far different. After all, "conservation of energy" simply means something *different* now than it did when written by Newton (Barrett, 2008). What, then, does it look like for one of *my* beliefs to be true? It stands to reason that at least some of the true propositions are as incomprehensible to human beings today as "Two elementary particles with half-integer spins cannot occupy the same quantum state at the same time" would have been to Isaac Newton.

Put another way, one might worry that the model of the world we use today is vastly different than the model in use at the end of inquiry, *M*, whose sentences are a complete representation of the real. The differences between the models might in fact be so pervasive that *nothing* we say today is true, and therefore *nothing* we say we know today is actually known. Due to the paucity and inaccuracy of our best model of the world (or rather, the best constructible model of the world given contemporary scientific, moral, philosophical, etc. concepts), none of our beliefs are strictly-speaking true. In the best-case scenario, where some one of our concepts is close enough to the world that it suffers no revision in future theories, it is hard to see how our arriving at our current understanding – as opposed to a different understanding that is compatible with our current model of the world, but woefully wrong according to *M* – is anything but extremely lucky. Even if we have *true* beliefs, the fact that they

are true in the first place can only be a matter of luck when we recognize the rampant inaccuracy of our best science.

It is perhaps unfortunate that the preceding objection is accurate. Our current best scientific theories, when taken together, are false. The standard interpretation of quantum mechanics is incompatible with general relativity (Barrett, 2003). Much worse, we have no idea how we can fix their incompatibility. In a similar way, quantum electrodynamics (and quantum field theory in general) involves procedures that are theoretically unjustified: to calculate the charge of particles, we must apply a technique known as *renormalization*, in which an infinite sum is subtracted from a divergent integral (an integral whose solution is infinite) in order to obtain a finite result. This procedure, though enormously empirically successful in that it enables us to obtain correct measurements and make correct predictions, is a result of quantum electrodynamics being what Cao and Schweber (1993) call *conceptually unstable.* A conceptually unstable theory is a theory that, though (possibly) consistent in that it does not make contradictory predictions, is undefined and undecidable in some portion of the domain to which it applies. Although we can, in the case of quantum electrodynamics, apply a renormalization procedure to obtain finite (and extremely accurate) results, the procedure itself is not justified by the mathematical foundation of the theory, but rather by the fact that the theory (i) Gives wildly inaccurate (infinite) results and (ii) Gives accurate results when renormalized. The procedure is illegitimate from a purely mathematical point of view, which lead to Dirac's lament that the procedure was "arbitrary" (1963, p. 53) and "not sensible mathematics" (quoted in Kragh, 1990, p. 184) and Feynman's proclamation that renormalization was a "dippy process" tantamount to "hocus-pocus."(2006, p. 128).

Our best scientific theories, quantum mechanics and special relativity, *work*. They enable us to make extremely accurate predictions, build global-positioning systems, etc. And they contain their own basic terms, to which we must be ontologically committed (so we must believe in electrons and photons, numbers and fermions). But quantum mechanics is not the full story: it involves a procedure motivated not by *the theory itself*, but by the fact that the theory predicts infinite results to problems whose solutions are verifiably finite. Worse, quantum mechanics contradicts relativity: *even if* we were able to modify the theory such that renormalization procedures were unnecessary (that is, such that they were *theoretically* as opposed to *empirically* motivated), it would still be the case that the conjunction of our two best theories is false. Not only do we *know* that our best scientific theories are false, we do not know *how* they are false. That is, we do not know what we could do to improve them. If we knew how our theories could be improved, we would improve them; the improved versions would immediately become the referent of "our best theories."

So we do not know how things could "get better." When we reflect upon the nature of scientific revolution, in which the basic terms of a theory are radically transformed in meaning, we have good reason to believe that what we mean now by e.g. "force" is radically different from what "force" will mean at the end of inquiry. Although we can currently understand how Netwonian mechanics were correct – and in what sense they are incorrect – by noting the ways in which relativistic mechanics recapture Newtonian predictions, the Newtonian scientist could not himself articulate ways to improve classical mechanics (though he could certainly recognize his model's failure to predict e.g. the precession of Mercury's perihelion, as did Le Verrier in 1859). In Ptolemaic terms, he could add epicycles to a planet's orbit about Earth, but could not see that the problem was not a matter of epicycles, but of the falsity of geocentrism.

If we start with our beliefs as a model, it is very difficult to see how any of them could be *true* in the fullest sense of the phrase described in Chapter 1. The conceptual structure that we marshal in propositionally articulating something as simple as "Amos is a dog" is radically incompatible with the conceptual structure that is *true*. While it is plausible that proper names might be fairly easily translatable from our current conceptual scheme to a true conceptual scheme (at least if we believe in direct reference), at the very least the term "dog" will mean something entirely different and – much worse – what we today mean by "dog" will very likely turn out to have no correlate in the external world. Of course, we cannot say how we are currently going wrong – if we could, we would self-correct.

The potential skeptic worries that the model *M* that satisfies true sentences looks nothing like our current model of the world. The descriptive content of our variables and predicates of our current language is quite different from the descriptive content of the variables and predicates of the language of the end of inquiry. In short, our model is not *M*, and so it is difficult to see in what sense *M* could satisfy sentences of *our* language. Moreover, the potential skeptic appears to have a strong inductive case behind his argument: scientific inquiry is not forward-looking, in that it cannot see how current theories are incorrect and move towards correction. Relativistic mechanics cannot be articulated in the language of Newtonian mechanics, and quantum mechanics cannot be articulated in the language of relativistic mechanics. There are myriad examples of radical conceptual divergence from one theory to the next, and we must agree with the skeptic that we know not *how* our conceptual shortcomings can be remedied, and so it begins to look like local true belief is hopeless.

Although the task of local true belief is difficult, it is not impossible. There are two similarly-stated but deeply different problems. The first problem is that many terms that we

generally recognize to be scientific in nature might turn out to not have referents in the real world and therefore to be false. Thus it might be the case that "W and Z gauge bosons have mass" is false because gauge bosons do not exist as currently conceived by our best particle physics. The second problem is that sentences that are not overtly technical or scientific might turn out to be false because the underlying conceptual schema in which they are articulated does not represent reality. For example, we might think that "Amos is a dog" is liable to turn out to be false because we currently think of dogs (along with all other material things) as composed of atoms, and it might so happen that our current concept of "atom" lacks any corresponding entity in the real world.

In the first class of sentences, we simply learn that an entity to which a theory is committed does not exist, full stop. In the sciences, we found out that phlogiston and the luminiferous aether simply are not part of the world. In the realm of everyday beliefs, a child might learn that "Santa Claus has a white beard" is false when she learns that Santa Claus is not a part of the mind-independent world (depending, of course, upon her philosophy of language). These are the instances in which we generally feel comfortable in saying that the relevant sentence is false. "The luminiferous aether is the medium of propagation for light" is false because there is no such thing as the luminiferous aether. This is not inherently problematic, as it is entirely appropriate that sentences stating the existence of fictional entities count as false.

On the other hand are the cases in which a sentence is rendered false because it employs concepts that either have no corresponding entity in the world, or the concepts it employs are dependent upon other concepts that do not refer. For example, 17th century chemists widely believed that air – the breathable fluid that "surrounds the whole surface of the terrestrial globe" (Macquer, p. 2) – was both homogenous and elemental. Thus one might take the 17th century

chemist (or layperson with some knowledge of then-contemporary science) to be saying something false when he says "The air is clear today" because, strictly speaking, what he means by "air" does not exist. Similarly, we might take the contemporary young Earth creationist to say something that is false when he says "Elephants are intelligent animals" because by "elephant" he means a creature that came into being roughly six thousand years ago in its present form. Intuitively, though, these sentences are true or at least can be true, regardless of what the speaker means by the terms "air" and "elephant."

If we admit that the 17th century chemist and the young Earth creationist say false things when make relatively benign comments about the external world, then the case for contemporary science and contemporary believers becomes quite bleak. Since we know that our current best science is false when taken as a whole, and since we expect that it will improve (though we cannot know how), then it appears that basic claims about the world are false. It is easy to identify how the young Earth creationist's statement that elephants are intelligent is false – it relies on a cosmology and conception of biology that are fundamentally at odds with the way the world is. While it is not so apparent just *how* "Elephants are intelligent" is false when said by a contemporary zoologist whose beliefs conform with contemporary science, it seems rather likely that what she means by "elephants" or "intelligent" (or both) will be superseded by future scientific progress in the fields of the evolutionary history of pachyderms and animal cognition.

To meet the above challenge, we must characterize a sense in which "Elephants are intelligent" is true *enough*. When its conceptual contents are fully articulated, "Elephants are intelligent" is probably strictly-speaking false, but there is (we hope) a sense in which it is translatable to a true sentence, a sense by which it is true, even if that truth is not the full-stop

truth of the conceptual schema at the end of inquiry. Articulating the sense in which our current sentences can be true will rely on what I call *charitable retrospective translation.*

While we cannot know how our best model of the world will improve, or in what sense it is in error, we *can* know how *previous* theories were in error. It is part and parcel of scientific progress that new theories – even when their vocabulary is radically different from that of the supplanted theory – can describe the error of previous theories. It is precisely this backwards-looking description of error and success that allows us to say that relativistic mechanics is a theory *of mechanics* – we recognize that *something* is preserved from the Newtonian formulation of mechanics, and that many of the sentences satisfied by the Newtonian model of the world are, when suitably translated, satisfied by the new, relativistic model of the world. It is the ability of relativistic mechanics to subsume Newtonian mechanics as a special case, describe Newtonian terms in relativistic terms, and point out how (some) Newtonian sentences were false that signals to speakers of the language that we are, in fact, talking about the same thing. While the Newtonian model has no room for relativistic concepts, the relativistic model *does* account for Newtonian concepts and vocabulary. From the standpoint of later theories, we can see in what sense supplanted theories were correct, and in what sense they were incorrect. Generally speaking, there are a few aspects of theories that are preserved to a greater or lesser degree in particular cases. On the short list of candidates for theoretical preservation are empirical predictions, structural and relational properties of the model, and existential claims regarding central entities of the theory (Barrett, 2008, p. 217-218). And we can identify such properties across various theories. From the standpoint of a current theory, though, we are in no position to state precisely what future theories will preserve.

The process of recapturing previous theories – charitable retrospective translation - must account for the empirical successes of previous theories. In most instances, at least some of the ontological commitments – and the vocabulary in which they are expressed – are maintained if for no other reason than to indicate that the subject matter of the two theories is one and the same phenomenon or group of phenomena. This process can only be backward-looking, or retrospective, because it is only from the position of our best theory that we can see how that theory reduced or removed error in the descriptive content of previous theories.[22] The process must indeed be charitable, because strictly-speaking, the sentences of the previous theory were false: they did not successfully refer to objects, relations, and so on in the external mind-independent world.[23]

To return to the case of the elephant, the zoologist can describe how the creationist's assertion is false (or, at least, how the current theory eliminates descriptive error). But in insisting that the creationist's sentence is false, it seems that we are making a mistake: although he is certainly incorrect that there exists an animal created for the pleasure of mankind six thousand years ago, there really *are* elephants. The descriptive content of "elephant" seems to be quite beside the point when we consider that both the creationist and the zoologist are talking about *elephants*, those large quadrupeds with prehensile trunks that we enjoy watching at the zoo. *Those* things certainly exist, regardless of what an individual thinks about their history,

---

[22] It is a simplification to speak of "the" way that an old theory reduces error – there are likely many competing and irreconcilable metrics by which a new theory can eliminate descriptive error, and therefore there are likely many competing and irreconcilable ways that a local proposition can be true to some degree or another. In what follows, I speak as if it is given that there is a single measure of truth for any proposition, but my understanding of truth and inquiry neither requires nor implies such a single measure, though there being such a single measure is a live option.
[23] It could, of course, be argued that they instead refer to *fictions*. But I am assuming that "Ahab is a man" and "Ahab is a fictional man" have different truth-values: the former purports to describe the real world, while the latter proposes only to describe a fictional world. Insofar as "Ahab is a man" purports to say that Ahab is flesh and blood like the rest of is, it is surely false.

evolutionary or otherwise. When speaking of everyday observable phenomena like elephants and air, we allow vagueness in the descriptive content of conceptual terms insofar as there is a readily-observable phenomenon that ensures our sentences refer. Whatever necessary and sufficient conditions we might use for attributing elephant-hood fall by the wayside in the face of Mrs. Jumbo. We do not need to worry that the creationist and the zoologist are talking past one another in discussing elephants because they both agree that they're talking about *those things there*. Similarly, we allow vagueness in the descriptive content of the modern-era chemist: what he means by "air" isn't *just* something elemental and homogenous, but also that stuff we breathe. While there is no single element that composes the gas around us, we can recognize that the air we speak of is the same air that the chemist spoke of, and that while he was surely wrong the descriptive content he attributed to the word "air," he at least managed to refer to the very same stuff.

The vagueness we presume in the descriptive content of everyday terms when in conversation with those sharing radically divergent belief-sets is not so different from the way that we can identify truth in past scientific theories. Just as in everyday conversation "elephant" becomes shorthand for "that large grey, betrunked animal and those like it," "air" becomes "that breathable gas around us" and "heat" becomes "whatever force makes things feel warm, and whose absence makes them feel cool." That previous theories variously understood heat as the presence of fire particles in a substance and as identical with the transferable fluid of caloric instead of as molecular motion does not imply that all sentences involving temperature of the past theories were false: because Macquer, Lavoisier, and contemporary scientists can all agree that heat causes the observable sensation of warmth, we are relatively comfortable in stating that all the preceding theories are indeed theories of *heat*. Just as the zoologist takes the creationist to

indeed be talking about elephants (despite being rather wrong about them), the contemporary chemist sees (or can see) her thermodynamics as a refinement of the previous theories.

It is of vital importance that we understand the notion of progress – both in the sciences and in the development of individual belief networks – as retrospective. As Peirce was fond of saying, we take each of our beliefs to be true, and that is what it *is* for them to be beliefs and not doubts. The intimate connection between belief and judgment is precisely that which gives us pause when presented with an example of Moore's paradox, such as "Elephants are intelligent, but I don't believe it." In the first part of the sentence we judge that something is the case. This judging is an action, the action is guided by rules we call beliefs. Barring obvious explanations like lying, it is hard to understand how one could manifest an action according to a proposition *p*, and in the same sentence manifest an action according to its opposite. That someone would sincerely judge an instance of Moore's paradox to be true is literally beyond belief. While we take our beliefs to individually be true, we generally expect that some of them will turn out to be false. When a rule for our action turns out to have unexpected consequences, we revise our rules in such a way that we can better predict the outcomes of, and therefore be successful in, our actions. After changing a rule, we can see how our previous belief was mistaken. If nothing else, we can see that if the world acted according to belief that *p*, *q* would follow. I acted *p*-wise with the expected outcome that *q*. When *q* does not follow, I learn something: that *p* does not cause (or otherwise imply) *q*. From the standpoint of my current belief system, I can describe how my previous beliefs were incorrect, but I cannot articulate how I expect my beliefs to change in the future. I cannot begin to say what beliefs I currently have are incorrect without taking steps towards constructing a future set of beliefs that allows me to interact with the world more successfully than my current beliefs. Any claims as to the truth of my current beliefs are

claims as what of my current belief-set will be preserved in the future, despite radical conceptual changes. The fact that I have experienced radical conceptual shifts means that I can only say that, of my current beliefs, I expect that they will able to be described in the terms of whatever future beliefs I have, and that my future belief-set will be able to account for the local successes of my current beliefs. The preservation of my current empirical successes will allow me to see that the subject matter in question – largely, the external world – is the very same.

Take, for example, the case in which I learn that what I currently mean by the word "elephant" is successful insofar as it allows me to point to elephants and say "That is an elephant," but is a failure insofar as it pertains to the care and husbandry of elephants. Learning that I am woefully misguided about elephants (perhaps they are not, it turns out, the sorts of things that can fly by flapping their ears) has the potential to radically undermine the possibility of local truth in *all* elephant-related judgment, including my successful identification of elephants at the local zoological park. Although it would certainly be an option available to me to say that all my previous beliefs and judgments about elephants were incorrect (after all, nothing I pointed to could flap its ears to fly), I can preserve the truth (more precisely, the approximate truth) of my previous judgments and beliefs by allowing for vagueness in my previous terms, and accordingly expecting that my current beliefs are vague and open to revision. Thus locutions like "closer to the truth" or "approximate truth" become successive iterations in a process of eliminating descriptive errors. By saying precisely what errors our past belief-systems had, and how our current belief-system eliminates those errors, we characterize the sense in which our current theories are true, although this truth is certainly not full-blown descriptive truth of the end of inquiry.

The relevant criterion for local true belief, then, is not that the world is exactly as I expect it to be, but that my current beliefs will be articulable from the point of view of the truth. Truth is not something to which we match our current belief networks (and scientific theories), but rather a full descriptive characterization of all that exists. We expect that given the truth, our current beliefs will be able to be recaptured in such a way that we can see precisely in what ways we are correct, and in precisely what ways we are incorrect. We purchase commitment to the truth of local belief-claims at the expense of admitting imprecision into our local conceptual vocabulary.

7. Conclusion

A belief is a rule for action. We posit a system as having actions, intentions, and beliefs when we understand it as a minded, agential system. To the extent that a system's actions – including its thoughts - conform to a set of propositional rules, that system can be said to believe those propositions. To the extent that these beliefs are recognizable from the point of view of the full descriptive truth as primitive articulations of that same truth, they are true. While the truth of our local beliefs is not the full-blown truth of settled opinion at the end of inquiry, it is sufficiently distinct from falsity that it can serve as a component of knowledge.

The process of eliminating local descriptive error in our beliefs – of recognizing failures of the rules guiding our actions and their revision – is inquiry. Our best cognitive science gives us a good model of belief generation. This model – coherence – is the subject of the next chapter.

CHAPTER 3: COHERENCE JUSTIFICATION

When S believes *p*, S acts as if *p* is the case. She expects the world to behave in a *p* sort of way, and she is surprised if she sees evidence that ~*p*. If, from the standpoint of the full descriptive truth at the end if inquiry, we would be able to recognize p as a primitive (in the developmental, not logical, sense) articulation of some portion of the full descriptive truth, we say that p was approximately true, or true enough. This character of "true enough" – which relies upon a charitable translation of contemporary propositions into the model that is the descriptive truth about the world – is what we count as a candidate for knowledge claims. When we say that a judgment that *p* is true, and the corresponding belief-that-*p*, of which the judgment-that-*p* is an instantiation, we commit ourselves to the preservation of *p* both in future conceptual schemata and to our *own* continued belief in *p*. But true local belief, however charitably translated, is not sufficient to grant *knowledge*. For an agent to *know* p, as opposed to merely having a true belief that *p*, she must also be *justified* in believing *p*. This chapter explains synchronic justification as a coherence relation among beliefs.

1. The primary bearer of justification

There are two important senses of justification available to us; which we use depends on what we take to be the primary bearer of justification. On the one hand, we could take the justification of individual beliefs to be the primitive understanding of "justification." In this case, the justification of a network of beliefs could be an entirely different sort of thing (perhaps call it "understanding"), or network-justification could be derived from the justification or lack thereof of that network's individual beliefs. On this understanding, a particular belief *B* is

justified (or not) by the set of epistemically prior beliefs *S* from which *B* is inferred. If *S* (or some subset of *S*) is justified, then *B* is justified. In this sense, justification "flows" from *S* to *B*, but not from *B* to *S*. If justification were to flow (probably through intermediate beliefs) from *B* to *S*, we would call that "circular argumentation." On the other hand, we could take a network of beliefs as the primary bearer of justification. On this understanding, the justification of a particular belief *B* depends upon the justification of the belief network *N* of which it is a part and any relevant facts about *B*'s situation in *N*. Some of the reasons we have for *B* might – in the long run – be influenced by the very fact that *B* is in our belief-set. On this view, justification is a property of a network of beliefs. The reliability, veridicality, *et cetera* of the belief network determine (in part or in whole) the justification of any particular belief. This section argues that the second, network-first concept of justification is the one that we should use.

Although belief-first justification tends to be assumed by most, there is good evidence that we should approach justification as a systemic property of a network of beliefs. The most historically obvious advantage that network-first justification has is its solution to the so-called regress problem. On a traditional understanding of justification, we are justified in believing a proposition when we have a reason for believing it. We also consider that if the reason in question is to grant justification, it must itself be justified. An infinite regress of reasons and reasons-for-reasons is easily generated. To get around this regress, foundationalists argue that some sort of reason is self-justifying (though perhaps this justification is defeasible). Descartes thought that clear and distinct perceptions, which were indubitable, were a secure foundation for building knowledge. Although the Cartesian epistemic program does not have many adherents these days, more recent philosophers have made quite some headway in providing candidates for self-justified beliefs (see, for example, Huemer 2001, Pryor 2000, and of course Moore 1959a

and 1959b). Although it is likely possible to construct a conceptual system that allows us to make sense of beliefs where the chain of justification may end (or begin), network-first justification allows us to simply sidestep the issue: it is just a mistake to expect the flow of justification of beliefs to never double back on itself. Justification is not from one belief to the next, like links in a chain. Instead, it is a matter of the relations among many beliefs that connect to one another in various ways, like the vertices in a bidirected graph.[1,2]

In abandoning a linear conception of justification, we also abandon the idea that justification only flows in one direction. Network-first understandings of justification hold that the justification of a particular belief is derived from the justification of the entire network (and, perhaps, particular properties relevant to the belief in question's situation in that network). Thus there is no single one-way relation between a particular belief and the belief(s) that constitute its justification that can be isolated from the rest of the network. It is precisely the holistic integrated character of the network – and its adaptation to the inputs constituted by successful and unsuccessful action – that allow the network itself to be justified. To isolate a single belief (or set of beliefs) as *the* reason (or reasons) ignores that our judgments and the evidence that supports them tend to be accepted or rejected *together*.

We have good intuitive reasons to suppose that justification is not a wholly linear process. Consider the general proposition that things are more or less as they seem. On a linear model of justification, belief in this could justify my belief that there is coffee in my mug (because there appears to be coffee in my mug), that there is a lamp on my desk (because there

---

[1] Some contemporary accounts replace or supplement justification with some other property, like reliability or truth-tracking, in their account of knowledge. This will be discussed in Chapter 5.

[2] The graph must be bidirected to allow that, for beliefs b_1 and b_2, justification could be conferred upon b_2 by b_1 (or vice versa), b_1 and b_2 could be mutually supportive, or justification could be blocked from flowing between them (as in the case that $b_1 = \sim b_2$).

appears to be a lamp on my desk), and so on for an arbitrary number of sense perceptions. A strictly linear understanding of justification would have it that the original proposition – that things are more or less as they seem – would require justification of some other sort (or is self-justifying, *a la* Huemer, Pryor, and Moore). But what proposition could support this (alternatively, who determines self-evidence?) Intuitively, the (general and defeasible) reliability of my senses is supported by the fact that things indeed *are* as they seem, at least in the focal cases of medium-sized objects. My success in picking out the objects necessary for my day-to-day life and the relative absence of relevant failures certainly seems to indicate that my sensory apparatuses and the data they present to my consciousness come as a package, as an interrelated network of propositions that are mutually supportive. The empirical success of my interaction with the world supports the proposition that the world is generally as it appears to be, and the world's generally being as it appears to be supports my acceptance of e.g. visual stimuli in the absence of reasons to doubt their veridicality.

There is also good empirical evidence that network-first understandings of justification are superior. If we assume that we are indeed sometimes justified in our beliefs, then understanding how humans do indeed reason should provide some insight into what makes a belief acceptable for human beings. Read, Snow, and Simon's 2003 study showed that not only do our evaluations of evidence influence the judgments we make based on that evidence, but that the judgments we make influence how we evaluate the evidence upon which those judgments are based. As a brute matter of fact, humans reason in such a way that there is a two-way flow of influence between evaluations of evidence and judgments based upon that evidence. A linear understanding of justification rejects this bidirectionality as illegitimate. Insofar as humans – who as a matter of biological fact exhibit bidirectional influence between evidence evaluation

and judgment – sometimes have justified beliefs, there must be a sense of justification that allows for such bidirectional doxastic influence. The holistic understanding of justification, wherein it is belief *networks* that are primarily justified and the justification of individual beliefs is derivative from network-justification – is (or can be) precisely that sense.

2. Coherence as Justification

When we speak of the justification of a belief network, we traditionally do so in terms of that network's coherence. That is, the coherence measure of a belief network just *is* the measure of how these beliefs interrelate, influence one another, and otherwise "hang together." Currently, the best model we have of coherence is that proposed by Paul Thagard. Thagard's work on coherence as constraint-satisfaction has successfully modeled both evolutionary and revolutionary scientific advances, including the transition of gastroenterologists from understanding peptic ulcers as caused by stress and diet to the contemporary belief that they are usually caused by the bacteria *Helicobacter pylori* (2002) and Lavoisier's conceptual revolution that abandoned phlogiston in favor of oxygen and caloric (1992, p. 82). Recent literature in psychology and the cognitive sciences suggests that understanding coherence as constraint satisfaction *a la* Thagard models not only scientific practice proper, but also run-of-the-mill human intuitive decision-making (Thagard, 2006), social reasoning (Read Snow & Simon, 2003), including reasoning about cause and effect (Read & Marcus-Newhall, 1993), change in attitude (Spellman, Ullman, & Holyoak, 1993), analogical reasoning (Spellman & Holyoak, 1992), and even reasoning about abstract rules (Holyoak & Simon, 1999; Simon, Pham, Le, & Holyoak, 2001). Constraint-satisfaction models have successfully predicted two-way flow of

justification, such that the conclusion one reaches influences the way that one perceives future evidence (Read Snow & Simon, 2003). Although the notion that we interpret data through the lens of our current beliefs is commonsensical, Bayesian, subjective expected utility, and information-integration theoretic models all predict that evidence influences the final conclusion one draws, but that the conclusion to which we are building has no influence on the way we interpret the evidence even as we are forming our conclusion. While constraint-satisfaction models will almost certainly be superseded by future models of judgment and cognition, they are currently the state of the art in experimentally predicting how individual human agents make decisions across many domains. If empirical humans are indeed *ever* justified (in a non-lucky way) in coming to conclusions, the best model that we have of that decision-making process is one of constraint-satisfaction.

The coherence relations that constitute simple synchronic justification are not just how humans *should* reason, but how we *do* reason. Our best model of human decision-making is a coherentist model of constraint satisfaction. Competing plans of action, goals, and so on are weighed against each other, and the most coherent plan is chosen and enacted. We act according to rules, and these rules are called beliefs. Since actions are the evidence from which we draw conclusions about beliefs, taking action just *is* taking up a belief (generating a new rule for action) or reaffirming an old belief (acting by a previously-established rule). Thus a descriptive account of how we act is by extension an account of how we believe.

This chapter will offer a constraint-satisfaction model of belief formation and justification. A network of potential beliefs is entertained (though unconsciously) and the most coherent belief-set is brought to life through action. This is the theory originally pioneered by

Thagard and others, and my work is an extension of theirs. My contribution comes in the form of a new coherence measure, whose benefits I will argue.

Importantly, this chapter describes how we can understand actions and their respective rules. It will *not* provide an answer as to what particular coherence measure counts as "coherent enough" for a belief to count as justified and therefore as knowledge. To the contrary, I take it that while coherent belief-sets are epistemically *better* belief-sets, mere coherence of beliefs, even when it is coherence of *true* beliefs, is not enough to grant knowledge. While I take it that a belief must increase the coherence of a belief network for it to be a candidate for knowledge (and that this constraint is only rarely if ever violated by individual believers), there are further diachronic constraints on knowledge, or more precisely, constraints on *knowledge-systems*, or *knowers*. These methodological constraints – what it takes to be a knower – constitute *diachronic* justification and will be the focus of the next chapter. For now, I will discuss *synchronic* justification, which I take to be identical with coherence.

The basis of my constraint-satisfaction model is that of pairwise dispositions to believe together and to believe exactly one of a pair. Before constructing the system from dispositions, it will be useful to say a bit about their nature.

3. On Dispositions

A disposition is structurally quite similar to a belief, though without necessitating any particular interpretative framework. Where attributing beliefs to a system implies that we also attribute actions, intentions, and a mind to that system, attributing a disposition to that system implies a commitment only to regularity in systemic outputs. We say that a system is disposed to

a particular output under a particular circumstance just in case, given that particular circumstance, the system generates that particular output.[3] This rather basic analysis was first suggested by Lewis (1997) and further developed and defended by Choi (2006, 2009, 2010). The simple conditional analysis of dispositions states:

(SCA) A system *S* has disposition *d* to produce output *o* under conditions *C* if and only if: If *S* were to be placed in *C*, *S* would produce *o*.

There are many proposed counterexamples to SCA (see Lewis, 1997; Lee, 2010; Lee, 2011). All, I believe, rest on a fundamental mistake regarding the subjunctive. I will argue that SCA is a fine analysis of dispositions, but that it is incumbent upon the individual using it to recognize how little is actually said when we attribute dispositions to a system. It is precisely the ontological commitment to mindedness and rules of thought (i.e. rationality) that make attributing the particular type of disposition "belief" to a system such an interesting and rich decision; simple statements of dispositions are, like statements of belief, statements as to expectations about a systems' outputs that are ultimately falsifiable and subject to *modus tollens* denials.

There are, generally speaking, two base sorts of counterexamples to SCA. In one case, we would generally say that *S* is disposed to *o* in *C*, but conditions obtain such that *C* and *S* does not, in fact, *o*. In the other cases, we would generally say that *S* is *not* disposed to *o* in *C* (or is disposed to ~*o*), but *C* obtains and *S* unexpectedly does indeed *o*. The former cases I will call *blocking* cases (since circumstances happen that *block* an otherwise-disposed system from acting

[3] Obviously, beliefs then are a special kind of disposition: they are the dispositions of minded systems to act in certain ways – that is, to generate outputs according to certain intentions.

according to its disposition), and the latter I will call *reproductive* cases (because circumstances occur that *reproduce* the effects of a disposition in a system that does not have that disposition).

As an example of a case of blocking, consider Martin's Holy Glass (1994): there is a piece of glass. We normally would say that the glass is fragile, because fragile means "would break if struck." But there is an omnipotent being that very much likes this glass, and ensures that if it were to ever be struck, it would melt (and so not break) and then immediately after being struck return to its previous form. Thus we would normally say the glass is fragile, but whenever it would be struck, a wizard (or deity) intervenes to make it such that the glass doesn't break. But then it goes back to its previous state.

As an example of a reproductive case, consider Martin's Living Dead Wire (1994): Suppose that we say a wire has the disposition "live" if and only if when touched by a conductor, electricity flows from it to the conductor. Now consider that the wire is hooked up to a machine that instantaneously detects when the wire is touched by a conductor. When the machine does not detect the wire's touching a conductor, it does nothing and so the wire remains dead (or non-live). When the machine detects that the wire has been touched by a conductor, it makes it live.

In the Holy Glass case, the fragile disposition of the glass ensures that if it were to be struck, it would break: $A > C$. Then, strange things happen such that A, but C does not obtain (yet we still want to say that the glass was fragile). In the Living Dead Wire case, we want to say that wire is dead and not live, which is to say that it is not disposed to conduct electricity when touched by a conductor. But the machine ensures that in all scenarios where it is so touched, it conducts. So $A > C$, the purported definition of "dead" is "$A > \sim C$," and yet we still want to call the wire "dead."

In each of the above subjunctive conditionals, the antecedent *A* must be understood to include both a particular output of a system (breaking in one instance, conducting electricity in another) and the circumstances under which we expect that output. What does it mean, though, to say that *if* the antecedent *were* to occur, then the consequent *would* obtain? More particularly, I must say what sort of semantics that the subjunctive conditional > receives. After all, "$A > C$" on the lips of van Fraassen (1966) and Stalnaker (1980) mean quite different things. [4]

A counterfactual conditional is a sentence of the form

(1) If such-and-such were the case, then so-and-so would be the case

Intuitively, sentences like (1) are true just in case so-and-so is true whenever such-and-such obtains - talk of "if" and "then" naturally leads us to consider the material conditional. When considering the material conditional as a candidate for the analysis of counterfactual sentences, however, we are immediately stricken by the ways in which it does not fit: "such-and-such" is actually false (these are *counterfactual* conditionals, after all), and as such we need some notion about the ways that the world might be, about what would happen if such-and-such were in fact the case. We call these ways the world might be *possible worlds.* However we understand the notion of the possible world, it is obviously useful in service of giving of truth conditions to sentences like (1) above. Without the notion of the possible world, it is unclear that (1) could even *have* truth conditions. While the no-truth-value position is not unpopular,[5] common intuition seems to uphold that sentences sharing (1)'s form *can* have truth-values. Perhaps not

---

[4] Van Fraassen does not, of course, directly discuss the subjunctive conditional. A supervaluational semantics of the subjunctive follows fairly naturally from the discussion. The present section describes such a semantics.
[5] See Edgington (1995) for a compelling account in this avenue.

all instantiations of (1) are truth-apt, but surely it is absurd to say that the following are truth-valueless:

(1a) If *F* were the case, then *F* would be the case

(1b) If *F* were the case, then ~*F* would not be the case

(1c) If *F* were the case, then *G* or ~*G* would be the case

(1d) If *F* were the case, then *G* and ~*G* would *not* be the case

It appears, then, that we need *some* notion of possible worlds if we want an (satisfactory[6]) account of what it is for a sentence like (1) to be true. In short, it is generally accepted that possible worlds are the truthmakers for counterfactual sentences. I take it that it is this status as truthmaker which is most important to a possible world; my view amounts to allowing for the properties of possible worlds to be determined not by a metaphysical view into what they are, but by their functional role in a counterfactual semantics. Thus it is not the metaphysician who tells us facts about possible worlds, but rather the semanticists and logicians - possible worlds are defined as precisely what they must be in order to allow counterfactual sentences to receive a truth value, and nothing more. As such, their properties are not determined by metaphysics or ontology, but by the languages whose semantics call for them.

My understanding of possible worlds – and therefore the counterfactual conditional – is constructivist: possible worlds are model structures created by an individual. A possible world is a sort of fiction, generated in talking about counterfactual situations. One might think of possible worlds as stories or narratives; there is a possible world at which the events of *Moby*

[6] This is to say, not a view that simply ascribes all counterfactual conditionals truth or falsity and calls it a day.

Dick play out just as described in the work, and it is this possible world whose defining elements Melville recorded. The reader of *Moby Dick* constructs such a world, although it is likely that it will not be the exact world upon which the novel was based - after all, there is no recorded height for Starbuck, nor is there an age for the Manx sailor. Thus we see that the text of *Moby Dick* corresponds to a number of possible worlds, possible models in which the relations and objects share certain features (Ishmael survives the sinking of the *Pequod*; Queequegg, Daggoo, and Tashtego are harpooners), but others undefined in the text are allowed to vary (what blend of tobacco Stubb smokes in his pipe). My understanding of a counterfactual sentence is similarly loose: those features defined in the antecedent must obtain, but those left unarticulated are allowed to vary. By supervaluating over antecedent-satisfying model-completions (each of them constituting a possible world), it becomes the case that a counterfactual conditional sentence $A > C$ receives a truth value when said value is determined by logical laws, but also when we fully specify our antecedent assumptions such that all model-completions agree as to whether or not C. Thus we can know precisely what it takes for $A > C$ to be true, and can discuss our antecedent assumptions once they are clarified and brought to the forefront of discourse.

The above presents a fairly standard supervaluational understanding of counterfactuals; for the complete semantics see Appendix A. Importantly, the only sentences containing counterfactual conditionals A > C that receive truth values on this supervaluational semantics are those in which C is true in all models that include A. Thus if it is logically possible that $A \& \sim C$ and it is logically possible that $A \& C$, $A > C$ receives no truth value. We can ensure that $A > C$ is either true or false by including particular assumptions in the antecedent. Top candidates for these usually include our best sciences. Although it is logically possible for me to mix two atoms of sodium with a molecule of chlorine gas and fail to obtain two molecules of sodium chloride, our

understanding of chemistry holds it such that $2Na + Cl_2 \rightarrow 2NaCl$. Given that we really have *no idea* of what kind of world we'd be living in if $2Na + Cl_2$ *didn't* yield 2NaCl and we therefore could not say anything sensible about such a world (beyond, perhaps, that the tautologies are true, the contradictions false, and so on), it seems desirable to set the truth value of "If I were to mix two atoms of sodium with a molecule of chlorine gas, the reaction would yield two molecules of sodium chloride" to true. We do this by "building" the sentences of contemporary chemistry into the antecedent and thus forcing the consequent to follow logically from the antecedent (which, again, is specified as something like "contemporary chemistry + I combine two atoms of sodium and one molecule of chlorine gas").[7]

As with a material conditional or any valid argument, you cannot get more out of a counterfactual conditional's consequent than was built into its antecedent. For all counterfactual conditionals $A > C$ that receive truth values, C either always obtains when A, or C never obtains when A. The trick to unpacking true counterfactual conditionals is in fully specifying the antecedent that so necessitates the consequent.

Let us return to the oft-disputed dispositional term "fragile." We generally think that something has the disposition of fragility just in case it has the disposition to break when struck. One might take SCA to analyze fragility in something like the following way:

(SCA) A system S has the disposition d to produce output o under conditions C if and only if: If S were to be placed in C, S would produce o.

---

[7] Interestingly, we *cannot* sensibly "build" the entirety of our best science into the antecedents of counterfactual sentences because our current best science is contradictory. Whether a counterfactual scenario assumes the possibility of instantaneous events (e.g. the wave collapse) or denies it (as in relativity) determines the space of possibilities and the truth value of counterfactual conditionals.

and

(Fragile) *S* is *fragile* just in case it has the disposition to break when struck

so

(SCA_F) *S* is fragile just in case if it were to be struck, it would break.

As an example of a fragile system, consider a large sheet of plate glass. This will be our paradigm case of fragility. Of course, it is not true that in *all* cases that glass is struck, it breaks. If it were struck by a tiny, angry mouse it would not break (we assume that the mouse is not travelling at high speeds). If it were struck by a cotton ball it would not break. (SCA) is nice because it has built into it a set of conditions that specify when a particular output allows us to attribute a disposition to the system. The rather innocuous-looking phrase "conditions *C*" hides the deep complexity that SCA allows (and indeed demands). The conditions are those assumptions – including e.g. contemporary chemistry – that necessitate the truth of the consequent. In the case of fragility, the conditions under which breaking entails fragility include that the striking body strikes with a force that is within a certain range, that the elasticity of the striking body is within a certain range, and so on. The conditions clause must also specify that there are no confounding forces at play: for example, if a ceramic vase is soundly secured in packing material (or magically turns into rubber and then back again when struck), the fact that it does not break when I throw rocks it is rather irrelevant to whether or not the vase is fragile. The further we specify the conditions, the more obvious it is that the sentence $A > C$ means something like "If *A* were to happen and everything is set up such that *C* must follow, then *C* would happen." While this is a true sentence, it is not particularly informative.

The antecedent of a counterfactual conditional must entail the consequent if the consequent is not a tautology and the conditional is to be true. In the case of dispositions like

fragility, cases of finkish wires, wizards, and padded vases give us reason to question whether the subjunctive conditional is adequate for an explanation of dispositions. I think that it is. In each of these cases, strange occurrences purport to divorce the antecedent and the consequent. In every instance, the antecedent can be further refined, and the conditions under which it is appropriate to count output *o* as sufficient for attribution of disposition *d* further specified. But analysis must stop somewhere. That is, we must *eventually* assume a common ground. Usually, that common ground would be something like our folk physics (perhaps informed to a greater or lesser degree by contemporary science).[8]

Many of the counterexamples to SCA (Holy Glass in particular) amount to answering a statement that $A > C$ with "Yes, but what if A and $\sim C$?" Divine intervention, or magic, or a miraculous machine, or what-have-you are sketched to make denying C while maintaining A seem plausible, but there is not much that can be said about the general structure of these A & $\sim C$ worlds. After all, we have *no idea* what the world would be like if wizards existed, or if glass spontaneously turned to rubber when struck, or if machines could produce all the physical effects of a wire's being live but there is an independent fact of the matter such that the wire (despite all possible evidence to the contrary) is dead. Such narrative devices in thought experiments are like the old philosophical saw of the evil deceiver and external-world skepticism: they amount to a statement that everything *seems* the same, but is *really* different.

---

[8] It's worth noting that other popular views of dispositions – say as properties – must face the same difficulties as SCA-theories. Any theory is most interesting at its limits, which is to say that the real working-out of SCA-theory happens when we decide whether we want to build a lack of wizards into the antecedent of the conditional associated with the disposition *fragility*. As a pragmatist, I tend to think that not much hangs on such questions, but for traditional realists, these questions are quite pressing. And, indeed, where SCA-theorists would argue about whether the conditional used in analyzing "fragile" should exclude wizards, property-theorists will argue about the circumstances in which a system must produce a certain output (i.e. manifest a certain property) for that property to rightly be called a disposition.

I can make no sense of what it means for an object to necessarily *seem* one way while really *being* another in the permanent, epistemically inaccessible sense above. Of course, it is perfectly sensible for things to seem to be one way *to us*, but for them to be another. This is simply being wrong about the world, and it is no mystery. For us to be incorrect, though, there must be some sort of mismatch of our beliefs on the one hand and the way the world really is on the other. Our beliefs lead us to have certain expectations about our actions, and the world fails to conform to these expectations. Cases of wizards and wires and evil deceivers posit not just that we're *currently* wrong, but that once all the data is in, we'll *still* be deceived.

Of course, much of the contemporary literature does not take itself to be creating evil-deceiver style counterexamples. Rather, they take it that the various counterexamples constitute legitimate cases of something's intuitively *being fragile* (or *live*) while not conforming to a particular subjunctive conditional. For these cases to have any force, though, they must have import above and beyond the obvious point that a system can have (or lack) a disposition that a proper part of that system lacks (or has). It seems quite obvious that if we *know* that the wizard is reliably preventing the glass from breaking, or that a machine is making a wire appear to be live, we would easily attribute non-fragility or live-ness to the wizard-glass or machine-wire system, as the case may be. That we cannot easily attribute systemic properties to the proper parts of a system should not be surprising. As such, I take the more pressing point to be that wizards (or machines) might confound us in such a way as to permanently divorce human epistemic capabilities from reality. If we can account for the strange external condition – if we can isolate the subsystem (e.g. the wire or glass) in which we are interested and talk about it in

isolation of the confounding larger system, we can sensibly use SCA.[9] As such, if these objections are to have merit, the wizard, the machine, and so on must be impossible to detect, such that they form an impenetrable epistemic barrier between the systems being analyzed and ourselves. And *this* is the same sort of thing as Evil Deceiver-style skepticism.

Of course, some of the literature does indeed take the fact that a system (e.g. wizard-glass) and a proper part of that system (e.g. the glass alone) can have different dispositions to constitute a counterexample to SCA. The most cogent such essay is that by Michael Fara (2005). Although he forcefully argues against SCA (though he does not take into account the individuation of systems, i.e. that glass-systems and glass-and-wizard systems are different things), he aptly distills the counterarguments to SCA into a single case: it seems to make sense to say "The glass is disposed to break when struck" and for me to occasionally strike it (even firmly, say with a hammer), but on rare one-in-a-million cases the glass does not break. Sometimes things seem to do what they are disposed to avoid, and sometimes things fail to do that to which they are disposed. And *this* is why a supervaluational semantics of the counterfactual conditional is vital. Consider two of the most popular semantics for counterfactuals: those of Lewis and Stalnaker. On Lewisean semantics, a counterfactual sentence $A > C$ is true just in case in *all* A-worlds of a certain closeness (and closer), C. On Stalnakerian semantics, $A > C$ is true just in case in the *closest* possible A-world, C. Here, the closeness ordering of possible worlds gets in the way and we begin to run into trouble. Let's say that there really is a one-in-a-million chance that the glass does not break. Is that single $\sim C$ world further from the actual world than the C-worlds? To

[9] This sort of system-division also accounts for the usual need for dispositions of a system to be due to some intrinsic property of that system. Intrinsic properties are those properties that are essential to – constitutive of – a system's being *that particular* system. Intrinsic properties of a system S are the only properties that *can* cause any sort of output of S, as opposed to some other system of which S is a proper part.

what extent do the 999,999 *C*-worlds count as individual worlds, and not a single world? If the only question we are asking is whether the glass breaks, it isn't clear how worlds should be individuated (particularly if we are interested in avoiding question-begging). To bypass this problem, we turn to a supervaluational semantics of subjunctive conditionals, in which there *is no* closeness ordering working to determine the truth and falsity of sentences. Rather, the conditional is only true or false according to the logical properties of *A* and *C*. If a sentence like "If the glass were struck, it would break" is to have a truth value, the antecedent must specify *exactly* what is meant. If the antecedent specifies strikes of an appropriate force, appropriate hardness and elasticity of the striking surface, a contemporary understanding of physics and materials science, and so on, it will still of course be undetermined as to whether the glass would break when struck. Eventually, the antecedent must be specified in such a way that the rare one-in-a-million case where the glass does not break is ruled out. This could be done in a specific (if ultimately *ad hoc*) way, by particularly omitting the particular combination of force, direction, and so on that would fail to yield a break or it could do so in the more obvious (though imprecise) way, saying "Nothing particularly strange happens" or "Everything acts as normal."[10] Of course, all of these are just ways of building the logical necessity of the glass's breaking into the antecedent of the conditional. By the time we've fully specified the conditional's antecedent, it isn't surprising at all that we haven't said much by uttering the conditional. Rather, we

[10] A probabilistic semantics of counterfactuals could even build this into the meaning of >. I, however, prefer to separate metaphysics from logic as far as I am able. For a particularly cogent probabilistic semantics of counterfactuals, see Hannes Leitgeb's (2012a and 2012b).

stipulate that for the purposes of present discourse, whenever A happens (or would happen), C follows.[11]

To return to Fara's case, we want to say that glass has the disposition to break when struck, but also that occasionally we strike it and it does not break. On SCA, the fact that nothing particularly strange happens is built into the "conditions *C*" clause. We generally leave it unsaid, because of course dispositions are generally taken to be *usual* happenings, not *certain* happenings. By taking the counterfactual conditional in SCA to receive its truth value by supervaluation, this is brought to the forefront. The conditions clause of SCA simply *must* necessitate a particular output if the relevant conditional is to receive a truth value at all.
In using SCA for dispositions, one does not make a substantive claim about the nature of dispositions. When I say "The glass is disposed to break when struck," I mean "Under usual conditions, if I were to strike the glass, it would break." And the relevant "usual conditions" are quite simply that the glass does indeed break. The "usual conditions" restrict the completion of A-models to what "usually" happens. If *C* "usually" happens, then $A > C$. If ~*C* "usually" happens, then $\sim(A > C)$. If *C* happens sometimes (but not enough for us to say that it is particularly "usual" or "unusual") then $A > C$ receives no truth value. Thus when I say "Glass has the disposition to break when struck," I must be understood as limiting the conceptual space

---

[11] At first blush, this might make one question the wisdom of using a supervaluational semantics in the first place. After all, surely it is an advantage of Lewisean and Stalnakerian semantics that they can provide truth values to A > C without forcing us to beg the question as to whether or not C in A! This advantage is not so great. Both Lewisean and Stalnakerian semantics force the exact same question-begging nature of A > C when they make use of the *closeness ordering* among possible worlds. The relative distances of C-worlds and ~C-worlds from a given A-world wholly determine the truth-value of A > C for both Stalnakerian and Lewisean semantics. Supervaluational semantics is to be preferred precisely *because* it is transparent in how it treats counterfactual conditionals. Very few would be tempted to take "If A were the case, then A would be the case" as a significant analysis of anything; supervaluational semantics make it clear that "A > C" is not an *analysis* but rather a *stipulation*. It rules out the possibility of A & ~C for the space of discourse. Bringing up cases where A & ~C isn't a *response* to a statement that A > C – it is instead missing the point of what is said in stating "A > C."

to instances where the glass is struck and it breaks. Cases where the glass does not break – cases where there is a wizard, or deity, or the physical properties of my blow and the glass align just so – are deviant cases and are *ex hypothesi* barred from consideration. Counterfactual conditionals with truth values are not contingent. Instead, they are theorems derived from an axiomatic set of presuppositions and demarcate the space of logical discourse.

I will, then, take the following version of SCA to be sufficient for present purposes:[12]

SCA: S is disposed to believe (and disbelieve) *p* and *q* together iff upon considering only *p* and *q*:[13]

(i) If S were to believe *p*, then S would believe *q*
(ii) If S were to believe *q*, then S would believe *p*
(iii) If S were to believe ~*p*, then S would believe ~*q*
(iv) If S were to believe ~*q*, then S would believe ~*p*

S is disposed to believe exactly one of *p* and *q* iff upon considering only *p* and *q*:

(i) If S were to believe *p*, then S would believe ~*q*
(ii) If S were to believe *q*, then S would believe ~*p*

---

[12] My interpretation of SCA here takes a strong reading of what it is to disbelieve *p*, such that to disbelieve *p* is to believe ~*p*. A weaker reading would be to say that disbelieving *p* is simply lacking a belief that *p*. On the weaker reading, an agent could manifest a disposition to believe exactly one of *p* and *q* by having *only* a belief that *p* or a belief that *q*. The stronger reading requires that the disposition involves *believing p* and *rejecting q* (or *vice versa*). On the stronger reading, if one has a disposition to believe exactly one of *p* and ~*p*, then one also has a disposition to believe *p* and ~~*p* together. While I do not consider this multiplication of dispositions to be vicious (it is not, after all, a multiplication of epistemic states, but metaphysical states), I do note that it may in the future be desirable to weaken the interpretation such that a disposition to believe exactly one of *p* and *q* entails believing exactly one of *p* and *q* with no concern towards whether the agent believing *p* also believes ~*q* (and similarly for the disposition to believe and disbelieve together).

[13] I will refer to an agent's disposition to believe and disbelieve together henceforth simply as a disposition to believe together. In what follows, though, it is crucial that we remember that the fully articulated disposition is not only to *believe* together, but also to *disbelieve* together.

4. Coherence from Dispositions

From the understanding of dispositions and the foundational concepts of dispositions to believe together and to believe exactly one of a pair, we can develop a notion of pairwise coherence that we will later use to develop full-blown coherence of a belief network. S's belief *p* is pairwise-coherent with S's belief *q* just in case *S* is disposed to accept or reject *p* and *q* *together*. S's beliefs *p* and *q* incohere (are pairwise-incoherent) just in case S is disposed to believe exactly one of *p* and *q*. There is no relation of pairwise coherence among *p* and *q* if neither of these conditions obtain.

This understanding of pairwise coherence, being based upon brute dispositions, is almost completely unrestricted. Those suffering from what Frege diagnosed as a "hitherto unknown type of madness" (1997, p. 203) such that they are disposed to believe both *p* and ~*p* together are surely not the sorts of creatures we would call "coherent." Indeed, I think the mad individual's epistemic situation is far worse than incoherence: his set of beliefs does not admit of a coherence relation *at all*. We cannot say whether his beliefs are coherent or incoherent, because his beliefs do not conform to anything recognizable by us as a rational structure. As such, it seems desirable to restrict coherence such that the dispositions of an agent with a coherent set of beliefs satisfy several fundamental constraints that guarantee at least a minimal sort of rationality. For pairwise coherence to even be a cogent notion, our agent's dispositions to believe together (or exactly one of a pair) must satisfy at least the following:

(1) If S believes $p \,\&\, q$, then S has a disposition to believe p, q, and $p \,\&\, q$ together.
(2) If S believes p and $p \rightarrow q$, then S has a disposition to believe p, q, and $p \rightarrow q$ together.
(3) If S believes p, then S has a disposition to believe exactly one of p and $\sim p$.

Following Thagard's account of explanatory coherence, we can add a fourth condition:

(4) Explanans/explananda[14]

(i) If S believes that p explains q, S has a disposition to believe p and q together.

(ii) If S believes that $p_1 \ldots p_n$ explains q, S has a disposition to believe $p_1 \ldots p_n$ together.

(iii) The strength of S's disposition to believe $p_1 \ldots p_n$ together in (ii) is diminished in proportion to n.

The addition of the condition (4) allows us an at least rudimentary basis for agreement, disagreement, and so on. The constraints above are intuitive with regards to actual human agents: it is a brute fact that we try not to believe contradictions, we believe the conjuncts of (what we take to be) true conjunctions together, and we believe the consequent of a conditional if we believe that the conditional and the antecedent are true. Of course, these dispositions may not manifest.[15] They are often contradictory, and it would be impossible to satisfy them all. Indeed, it is trivial to construct a case in which both neural-net and exhaustive constraint-satisfaction algorithms accept contradictory statements (see Appendix B.1, Fallibility). Importantly, the contradictory beliefs at play in Fallibility are precisely those that we must believe. Of each belief, we take it to be true (necessarily so – if we did not take it to be true, it would be a doubt). We expect, though, that some of our beliefs are wrong, though we cannot say which. Although coherence does nothing to dissolve the contradiction of the paradox of the preface, the constraint-satisfaction model of it explains how we can take each of our beliefs to be individually true, while also believing that at least one of them is false.

---

[14] A plausible (and, I think, correct) account of explanation is given in Michael Strevens (2004) and (2008). On Strevens's understanding, if a set of statements $p_1 \ldots p_n$ is an explanation for event e, then $p_1 \ldots p_n$ are causal difference-makers for e. Thus in (4), (i) can be recast as something like (i*) "If S believes that event e caused event f, then S has a disposition to believe that e occurred and that f occurred together." (ii*) would be something like "If S believes that events $e_1 \ldots e_n$ caused event f then…"

[15] This adds an interesting complication to the antecedent for an SCA-analysis: something like "In the absence of other dispositions…" It rapidly becomes apparent that the Simple Conditional Analysis is anything but simple.

The ability to rationally believe contradictory propositions immediately creates a tension with the understanding of contradictory beliefs explained in Chapter 2. In Chapter 2, I said that the best understanding of any set of beliefs is one in which the agent does not contradict him or herself, because a consistent theory is always better than an inconsistent one. Since beliefs are theoretical entities posited to make sense of the outputs of belief-systems, the best application of them to the outputs of a system is, *ceteris paribus*, an application which gives that system consistent rules. The tension between these positions reveals an important aspect of belief-attribution and justification: the belief-systems that we can reasonably posit to ourselves and others are, much as in the case of our best science, *not* the best belief-systems possible. In the light of the full descriptive truth, we will be able to give any system of outputs a consistent set of rules. Our position from the inside, though, does not have this feature. In the case of physics, the full descriptive truth will resolve the contradiction between quantum mechanics and relativistic mechanics, suitably reinterpreting the sentences of the mutually contradictory theories into a single consistent system. From the viewpoint of contemporary physics, though, the contradictions remain. The best interpretation of the data – the next generation of physical theories – is not available to us (and if it were, *it* would immediately become our best physical theory). Similarly, it is not always available to us precisely *how* we can make sense of systems with (apparently) contradictory beliefs. We can be confident that once all the data is in, a consistent theory will be available, but this optimism about future inquiry does not help us to resolve local cases of the Preface Paradox, the Lottery Paradox, and so on. Believing contradictory propositions, as when we take our beliefs to be individually true but also believe that we are fallible and have some false beliefs, is a step towards having a consistent set of beliefs. When we learn that one of our beliefs is false, we change our minds about it and

gradually move towards a consistent understanding of the world and the belief-systems (including ourselves) it contains. Indeed, recognizing the contradiction between our acceptance of each of our beliefs and acknowledgment of our own fallibility can lead us to reconsider our individual beliefs. If they are found wanting, we change our stance with respect to their propositions. If we are somehow able to test each of our beliefs and find no logical contradictions, no failures in action, and no occasion for doubt, then we can reject the proposition that we have some false beliefs.

Although it is difficult to imagine a situation in which an actual human being could reasonably reject his or her own fallibility, toy cases with restricted belief-systems give us reason to think that such fallibility could in principle be rejected at the end of inquiry. Imagine Silas, who is very cautious in his actions. He only acts once per day, and that action is always the same: he judges that he exists. After several months of this, Silas begins to act twice a day: he judges that he exists, and he judges that his previous judgment is true. Some time later, Silas adds a third judgment to his daily routine, that he is fallible and so at least one of his beliefs (besides the one corresponding to this judgment) is false. Presumably after enough trials, he can reject his third belief. After all, he judges that he exists, and he finds no evidence to contradict this opinion. He also judges that his previous judgment was true, and again he finds no evidence to contradict this. The only further candidate for evaluation is his belief that one of his other beliefs is false. He has every reason to doubt this belief, for it contradicts his *other* two beliefs, of which he is certain. Absent further judgments, beliefs, and occasions for doubt, Silas is perfectly justified in sitting in his room and affirming "I exist," and "My belief that I exist is true" to himself (see Appendix B.2).

Of course, the above example is simplified in the extreme. We could posit alternative beliefs and interpretative methods to Silas: we could say that he is afraid of leaving his room, or that he is not a belief-system at all, and is more akin to a computer program that spits out sentences every twenty-four hours. Doing *that*, though, simply reinterprets the given system in a new way. Given that a new interpretative system has new ontological commitments and predicts different actions, it is not at all surprising that altering the rules of action attributed to Silas would alter what we say about him. The point, though, is that given *these* interpretations of his actions, Silas is (and we are) safe to conclude that all of his beliefs aside from the fallibility-affirming belief are true and reject the fallibility clause necessary for the generation of the Preface Paradox.

Coherence is the measure of the degree to which a network of beliefs "hangs together." While this familiar gloss is perhaps unilluminating, I follow Thagard in formally defining coherence as degree of constraint satisfaction. The constraints to be satisfied are the positive and negative connections that structure an agent's set of beliefs: the dispositional connections discussed above. A belief network satisfies these constraints by affirming or rejecting both of positively connected beliefs and affirming exactly one of negatively connected beliefs (and denying or rejecting the other). The sum weight of connections satisfied in this manner, subtracting out the sum weight of unsatisfied connections, is the *coherence measure* of a belief network.[16]

My understanding of the coherence measure is not, of course, the only one available. Plausible formulations include the following (the version briefly advocated above is C_2):

---

[16] Denial can take at least two forms: in denying *p* we may affirm ~*p* or simply withhold our affirmation of *p*.

$C_1 =_{df} \Sigma_i \Sigma_j S(R_i, R_j)$	Sum of satisfied weights[17] in belief system R
$C_2 =_{df} \Sigma_i \Sigma_j S(R_i, R_j) - \Sigma_i \Sigma_j U(R_i, R_j)$	Sum of satisfied weights in R minus sum of unsatisfied weights in R
$C_3 =_{df} \frac{\Sigma i\ \Sigma j\ S(Ri,Rj)}{\Sigma i\ \Sigma j\ T(Ri,Rj)}$	Sum of satisfied weights in R divided by sum of weights in R
$C_4 =_{df} \frac{\Sigma i\ \Sigma j\ S(Ri,Rj) - \Sigma i\ \Sigma j\ U(Ri,Rj)}{T \Sigma i\ \Sigma j\ T(Ri,Rj)}$	Sum of satisfied weights in R minus sum of unsatisfied weights in R, divided by the total sum of weights in R

In the case of C_1, the sum of unsatisfied constraints plays no part at all: the only relevant measure of the degree of coherence of an agent's belief network is the sum of satisfied dispositional connections among her beliefs. This seems to benefit the agent who is particularly blind to the incompatibility of her own beliefs as contradictions can be safely ignored so long as she does not notice them. If coherence is to at all reflect epistemic merit, C_1 cannot be its measure.

Having jettisoned C_1 as a proper measure of coherence, we must decide amongst $C_2 – C_4$. The difference between C_3 and C_4 is the subtraction of unsatisfied weights in the case of C_4. Whether we choose C_3 or C_4, though, is unimportant: in the case of an ideal agent $\Sigma_i \Sigma_j U(R_i, R_j) = \Sigma_i \Sigma_j T(R_i, R_j) – \Sigma_i \Sigma_j S(R_i, R_j)$, and so $C_4 = 2C_3 – 1$.[18] One would be free, then, to choose between C_3 and C_4 based on whether the coherence measure is to range between zero and one (inclusive), or between negative one and one (inclusive). Understanding the relationship between C_3 and C_4, they can together be contrasted with C_2.

Coherence measure C_2 is sensitive to the size of the belief network *R*, while proportional measures of coherence (C_3 and C_4) imply a notion of maximal coherence. That is, when it comes

[17] I do not at this point have a strong conviction that connections in the formal model have diverse weights (although it is intuitively plausible that we can be more or less disposed to believe *p* and *q* together). While Thagard does not include weights in his model, adding them would not be particularly difficult.

[18] That is, for an agent who "notices" all dispositional connections, the total weight of her belief network is equal to the sum of the satisfied and unsatisfied weights in that network. Thus the unsatisfied weight may be replaced by the value of the total weight minus the satisfied weight, and we get the result that the value of the coherence measure given by definition C_4 is twice that given by C_3, minus one.

to C_3 and C_4, size does not matter. Instead, what is central to coherence is maximization of explanatory (and perhaps other) constraints among propositions. The base question is this: Is it *ceteris paribus* better to have many beliefs? This question seems difficult at first; its solution lies in considering edge cases.

Assume an epistemically perfect agent in a particularly lucky epistemic situation such that the sum weight of unsatisfied constraints in her belief network is zero. Under measures C_3 and C_4, her coherence measure is one. Under C_2, however, her coherence measure is the sum of satisfied dispositional connections among her beliefs. The question, then, is whether adding a further belief (assuming no unsatisfied constraints) makes her belief network *more* coherent, whether adding further supporting beliefs makes the belief network *more* justified. Similarly, assume an agent who is particularly bad at reasoning from premises including his beliefs to novel conclusions such that he manages to satisfy *none* of his dispositions to believe together and exactly one (of a pair). The proportional measures give him their particular minimum value (zero or negative one), while the non-proportional measure assigns as his coherence measure whatever is the sum of unsatisfied explanatory weights. If we can say that accepting a further belief which does not satisfy any constraints makes him *less* justified (that is, *less* coherent) and epistemically *worse off* than he was before, then we have reason to accept the non-proportional definition of coherence.

It seems that in both the above scenarios, the non-proportional view is favorable. Certainly addition of a belief with wholly satisfied (or unsatisfied) explanatory constraints makes an agent more epistemically well off (or worse off). In the case of epistemically normal individuals in standard situations where it is not the case that all explanatory constraints are recognized and satisfied, the acceptance of a belief which satisfies a greater constraint-weight

than it does not satisfy will increase the coherence measure, and the acceptance of a belief which leaves a greater weight of dispositional constraints unsatisfied than satisfied will decrease the coherence measure.[19] The decision as to whether we use a proportional or non-proportional coherence measure is decided by the cogency of the notion of maximal coherence. Because it seems that a particular belief network could always be more justified (that is, could better cohere or "hang together" more closely), the non-proportional definition C_2 is preferable as a measure of coherence and justification.

There is further reason to prefer a non-proportional account of the coherence measure. As I understand it, (rational) consciousness obeys certain constraints in accepting justified beliefs, and the real task of a theory of coherence is that of explaining these constraints. Formally, the relevant constraint is expressed as the affirmation of beliefs which do not reduce the belief network's coherence measure. The broader *process* of the justified acceptance of beliefs, however, must be the primary object of understanding and investigation. On a proportional account of the coherence measure, we notice that there is a certain maximal degree of coherence, and so might conclude that a coherence measure of one is "real" coherence, that justification amounts to having a maximally coherent belief network, and anything short of this is unjustified or incoherent. We might take implied constraint on belief acquisition – belief that which increases – to serve as the method by which we can attain a coherence measure of one. Such conflation of condition and procedural constraint is less tempting on a non-proportional view. If the untenability of maximal coherence serves to remind us that our primary goal lies in

[19] This result could then account for the intuitive understanding of a particular belief's being coherent: a belief is coherent (or incoherent) with a network of beliefs to the degree that it raises (or lowers) the coherence measure of that network. In the case that the belief's addition neither raises nor lowers the coherence measure of the system, that belief is independent of the system.

explaining how we can come to accept beliefs which are justified and may count as knowledge, then it seems we have good reason to understand the coherence measure non-proportionally, even if the reason is wholly a pragmatic one.

5. The coherence of subsets

Having settled on C_2, it is useful to discuss its limitations. Intuition leads us towards a coherence measure that does not admit of a maximal degree of coherence, but this implies that our coherence measure is sensitive to the number of elements in a subset. As such, it is useful to give an algorithm for comparing the coherence of mutually incompatible belief networks in a way that does not increase the coherence of a theory (that is, a subset of a network) because of the mere fact of its rejection of another theory (that is, a separate subset of the network).

In the general case, if a belief-network *N* entertains two mutually incompatible theories T_1 and T_2, exactly one of them will be accepted (modulo shared propositions). If we add to our list of dispositional constraints a consideration of competition:

(5) Competition*:* If S believes that *p* explains *r*, S believes that *q* explains *r*, and does not believe any of (i) *p* explains *q*, (ii) *q* explains *p*, (iii) *p* and *q* explain *r* together, then S is disposed to believe exactly one of *p* and *q*.

we can include mutually incompatible theories into a shared decision-space in order to see if we should accept, for example, a Copernican or Ptolemaic cosmology. Unfortunately, constraint-satisfaction metrics do not give us a way to directly compare the Copernican system with the Ptolemaic system when each is taken in isolation. This does not, however, mean that we cannot say that a Copernican system is more coherent. What we mean by "more coherent," though, is perhaps somewhat unintuitive: T_1's being more coherent than T_2 means that accepting T_1 and

rejecting T_2 maximizes the overall coherence measure of a network that contains both T_1 and T_2 as live but mutually incompatible options.

This understanding of "more coherent" only gets us so far, though. It remains the case that if T_1 and T_2 have no relationship at all – if they are, for example, Copernican astronomy and a caloric theory of combustion – we cannot say that the former is more coherent than the latter. Since the two theories do not purport to explain the same data, they are simply irrelevant to one another and so there is no need for negative constraints between them. As such, insofar as each of the isolates T_1 and T_2 increases *N*'s overall coherence measure (as of course each must, for propositions of T_1 and T_2 will be accepted or rejected precisely in the way that does indeed best satisfy the particular constraints upon them and thus so increase the measure), it will be accepted, thus increasing the sum of *N*'s satisfied constraints.

This limitation, of course, is a function of how we use coherence as reasoners: coherence is a constraint on acceptance, such that we accept or reject a proposition according to which position maximizes coherence. While coherence justifies us in the empirical sense of being the best explanation for how we do indeed reason, it alone cannot provide justification in the normative sense. After all, if we have a silly set of beliefs, we tend to accept other silly beliefs on the basis of their coherence with our priors.

6. Conclusion

Having settled on a coherence measure for a network of beliefs, we are now able to define the coherence of particular beliefs: a particular belief is *coherent with* a belief network just in case its acceptance raises the belief network's coherence measure. Coherence measures

how closely a network of beliefs "goes together" as a single doxastic unit, but a belief's being coherent (that is, coherent with a belief network) does not guarantee that the belief is likely to be true. A belief's coherence with a belief network is only as reliable as the belief network with which it coheres. Thus brute coherence is untenable as a final analysis of epistemic merit. If our network of beliefs is generally false, based on fantasy, or otherwise defective, then we cannot expect the coherence of a belief with the rest of our beliefs to make our justified belief more likely to be true than if it did not so cohere. Indeed, given an agent with a sufficiently silly, deluded, or otherwise defective set of background beliefs, it might be the case that *his* being justified in believing *p* is a good predictor that *p* is false, and that *his* rejection of *p* is good evidence that, *ceteris paribus* we should accept it. Thus if we take it as given that justified beliefs are more likely to be true than unjustified beliefs, then coherence cannot be the entire story with respect to justification. While an agent is (synchronically) justified in believing *p* if and only if *p* is coherent with her network of beliefs, there is a further sense of justification that coherence does not guarantee. *That* type of justification – that is fully truth-conducive and so on – is determined by particular features of the methodology used by particular believers in belief-acquisition, and is the subject of the next chapter.

CHAPTER 4: METHODOLOGICAL JUSTIFICATION

A particular belief is *coherent with* a belief network just in case its acceptance (as opposed to rejection) increases the coherence measure of a belief network. Coherence measures how closely a network of beliefs "goes together" as a single doxastic unit, but a belief's being coherent (that is, coherent *with* a belief network) does not guarantee that the belief is it all likely to be true. A belief's coherence with a belief network is only an indicator of that belief's truth if the belief network with which it coheres is generally veridical. If our network of beliefs is generally false, based on fantasy, or otherwise defective, then we cannot expect the coherence of a belief with the rest of our beliefs to make our justified belief more likely to be true than if it did not so cohere (see e.g. Shogenji, 2005; Shogenji, 2007; Schubert, 2012; Olsson, 2005). Indeed, given an agent with a sufficiently silly, deluded, or otherwise defective set of background beliefs, it might be the case that *his* being justified in believing *p* is a good predictor that *p* is false, and that *his* rejection of *p* is good evidence that, *ceteris paribus* we should accept it (that is, he might be *anti-reliable*). Thus if we begin an analysis of justification with the axiom that justified beliefs are more likely to be true than unjustified beliefs, then coherence cannot be the entire story with respect to justification. This chapter is an investigation into a second type of justification that I call *methodological* justification.

First, I will discuss fallibilism as an essential component of a truth-seeking method of inquiry. The second section develops and explains methodological justification as embeddability in a truth-seeking practice; Sections Three and Four develop a formal definition of methodological justification. Sections Five and Six test this new measure of justification against

familiar Gettier cases and apply it to a foundational debate in the epistemology of disagreement respectively.

1. Methods of inquiry

Truth is the system of propositions that fully captures reality. Because true beliefs really *do* conform to reality, our actions based on true beliefs are not liable to be frustrated. If an interlocutor disagrees with one of our true beliefs, the very state of the world weighs in in our favor. A truth-seeking method, then, must take reality to be the final arbiter in all circumstances. A truth-seeking method must seek in all instances to preserve the phenomena, to account for *reality* as opposed to fiction or inclination. This, after all, is merely what it means for that method to be *truth*- (as opposed to *fiction*-) seeking. In this section, I develop what it is for a method of inquiry to be truth-seeking, ultimately concluding that methodological fallibilism is the cornerstone of a method's truth-conducivity.

Peirce (1877) identifies several (global) methods of inquiry. The fixation of belief is necessarily concerned with synchronic coherence, of course; the methodological commitment at issue is what we take to be the final indicator of veracity. In the dogmatic mode, we assert something to ourselves and simply refuse to change our minds or entertain beliefs that might occasion a change of mind. This is the *method of tenacity* and it is the easiest method of belief-formation. I believe whatever I believe, and I continue to believe it because I believe it now. Unfortunately, this method of belief-fixation faces severe difficulty in the face of inputs that contradict its established beliefs. When something with external permanency – like reality – pushes against a tenaciously-formed belief set, the belief set cannot accommodate this new data.

Massive and systemic doubt – that uncomfortable state – is usually the result. Eventually, this method of belief breaks down in the presence of others with different beliefs and more effective actions. Sooner or later the tenacious believer is bound to recognize that other people aren't, in fact, evil or corrupt or insane, and will be stifled by the law or common social decency from expounding upon his ridiculous theories – tenaciously held beliefs are not particularly long-lived.

The second method of belief-formation is quite similar to the first. This method, the *method of authority*, relies not on an individual's tenaciously-formed beliefs, but upon a belief-set given by some authoritative body. The church or government or political party defines what it means to be part of a community, and membership in that community is defined by rigid and tenacious adherence to its beliefs. Non-believers can be shunned, and failing that "a general massacre of all who have not thought in a certain way has proved a very effective means of settling opinion in a country" (Peirce 1877, p. 13). While the method of authority is more effective than the method of tenacity, this is only due to its definitional locus of power and ability to either isolate itself from contravening input or form a community based around ignoring that input. The method of authority in effect makes all members of a community tenacious believers in the *same* set of beliefs, and so there is little room for disagreement. Despite its advantages over the method of tenacity, the method of authority is also doomed to fail. Eventually isolation will be impossible or undesirable to enforce, and contravening data will enter an agent's belief-network. The aforementioned massacre and excommunication can only go so far; eventually doubt will be resolved not with tenacious anger but with the rejection of dogma, and once this process begins it can be quite difficult to end.

The third method of inquiry is the *a priori* method. On this method, I form a theory to best explain the data at hand, and that is the end of theory-formation. Original theoretical

commitments are held as sacrosanct, and the theory can only be extended and given additional complexity to explain new and recalcitrant data. Kuhn (2012) calls this "normal science," and while it works quite well insofar as the underlying theoretical commitments are *true*, eventually a particular theory must give way in the face of recalcitrant data and the growing realization that in the face of contradictory evidence, our original theoretical commitments were matters of preference and historical inertia, as opposed to true propositions that predict and describe reality.

Where the method of tenacity's primary methodological commitment is "Believe what I believe," and the authoritarian method's primary commitment is "Believe what I am told," the *a priori* method starts from a reasonable foundation, but allows historical and theoretical inertia to prize particular theoretical commitment over explanatory utility. All the aforementioned methods take themselves to be truth-seeking: the individual in the thrall of the authoritative method of inquiry takes the testimony of the relevant authority to *be* the truth, the tenacious inquirer takes the fact that he believes a proposition to be sufficient to guarantee its veridicality, and the *a priori* theorist clings so desperately to his collapsing system precisely *because* he takes it to be true. Importantly, though, it is not the case that these methods are *actually* truth-seeking. If a method is to be *truth*-seeking, it must *actually* take conformation with observable reality to be its sole criterion of success. Peirce calls such a method the "method of scientific investigation" (1877, p. 19).

Practically speaking, a truth-seeking method of inquiry must be fallibilist with respect to the beliefs upon which it operates. It should be willing to jettison those beliefs that cause failed action and thus occasion doubt. Similarly, it must be *methodologically* fallibilist, in the sense that it must be willing to doubt and perhaps jettison those methodological commitments that past experience indicates lead it away from capturing reality. For example, parsimony is usually

considered a theoretical virtue, but its virtue is inferior to that of actually explaining the observational data: when physicists can only explain the outcome of their experiments by positing a new particle, parsimony must be violated to better explain reality. This is just to say that reality takes precedence over theory.

It is an interesting question whether methodological fallibilism itself is open to rejection in a truth-seeking method of inquiry. Of course, it must be on pain of contradiction, for methodological fallibilism must be able to be rejected to satisfy its own methodological constraint. This, however, does not diminish the fact that methodological fallibilism *really is* truth-seeking: while we should be *open* to rejecting it, it certainly *seems* that we will never have occasion to do so.[1]

We must be open, then, to rejecting methodological fallibilism. What would such a rejection look like? First and foremost, we must inquire as to *why* we would reject it. The obvious answer is that methodological fallibilism causes occasion for doubt, that actions based upon it fail to conform to reality, and so on. In short, methodological fallibilism would be rejected just in case it turns out to *actually be* a problem for a truth-seeking methodology. Its rejection would indicate that at least temporarily, it is not a good method for eliminating descriptive error. Actually *rejecting* methodological fallibilism, though, is to accept that a particular methodological commitment was fallible, and is in fact to verify it. If methodological fallibilism is false, then we must reject methodological fallibilism – and in doing so we affirm it.

---

[1] The same can be said of *any* methodological commitment. After all, if we thought that our methodological commitments failed to get us to the truth, we would doubt them. That I do not doubt that which I take to most reliably soothe doubt is tautologous. Methodological fallibility is in an interesting position, in that rejecting that it is truth-seeking verifies its truth-conducive nature.

Formally speaking, this is an instance of the Law of Clavius, also known as *consequentia mirabilis*.[2] It is just a specialized form of *reductio ad absurdum*: when we can derive a contradiction from A, we reject A; in the posited case of A, the derived ~A may be conjoined with it yielding the contradiction necessary for the *reductio*.[3] In the case where A is "~p," we can reject it and thus accept p. As such, we can be as confident in methodological fallibilism as we are in proof by contradiction. Assuming we are not intuitionist logicians, there is no higher honor, and on pain of contradiction, any actually truth-seeking method of belief-fixation must be methodologically fallibilist.[4]

2. Methodological justification as embeddability

Methodological fallibilism is a prerequisite for a belief-formation method's truth-conducivity and therefore its being justified. This section develops an understanding of methodological justification as embeddability in a self-correcting belief-formation process by examining several examples of Gettier cases.

For a belief to be methodologically justified, it must be formed in a fallibilist way that always takes reality as the final arbiter of a proposition's truth. As mentioned above, the method differentiated is global, not local. Any local method can be used in the service of any of the global methods identified. I could, for example, be a tenacious believer in exactly what I believe

---

[2] Bas van Fraassen (1968) made ample use of this law (though not by name) in his supervaluational semantics for free logic, where he defined presupposition as the semantic relation where in one sentence follows from both the truth and falsity of another, such that *A* presupposes *B* if and only if *B->A* and *~B->A*.
[3] Alternately, one can derive *~A* from *A* thus proving *A->~A*, convert to *~A* v *~A,* and finally conclude *~A*.
[4] Of course, just as with empirical adequacy and parsimony from Chapter 2, there is nothing to stop what counts as "methodological fallibilism" at the end of inquiry from being quite different from what I here mean by "methodological fallibilism."

right now, rejecting all evidence that contradicts my current beliefs. The beliefs I have, though, were formed in a myriad number of ways, some locally reliable and some not. If I believe tenaciously, I am not methodologically justified.[5] I could form my beliefs by pulling them from a hat, but believe them in such a way that I recognize that they are merely the best bet I have for action right now and are liable to steer me wrong, and am willing to revise them in the face of contravening evidence. In such a case, my local pull-from-a-hat method is extremely unreliable, but is part of a global method of belief-formation that is truth-seeking. Although Peirce's scientific method of inquiry is reliable *in the long run*, there is no guarantee that local methods used to form beliefs are similarly reliable. This is the difference between a (global) method's being truth-seeking and a local method's being reliable. It is only at the end of inquiry that using a truth-seeking method necessitates the reliability of our local methods. In the meantime, an appropriately truth-seeking method can be instantiated in any number of reliable, unreliable, and anti-reliable local methods. It is our reaction and openness to revision of our local beliefs and local belief-formation methods that determine methodological justification, not local reliability.

For a belief to be methodologically justified, it must be formed in accordance with a method that is truth-seeking. For a method to be truth-seeking, it must take reality to be the final arbiter and determinant of truth. Accordingly, a truth-seeking method must take its methodological commitments to be fallible (though of course it believes its current commitments to be the right commitments). Importantly, it will be rare that we can identify ourselves as

[5] For the most part. As we will see below, appropriate fallibilist belief eventually hits upon the truth, at which point it is reasonable to be quite certain of one's beliefs. It must be stressed, though, that this possibility is solely conceptual. We never *will* be at the end of inquiry with full access to the truth, and so it will never *actually* be appropriate to believe tenaciously.

failing this criteria. We can attempt to diagnose others as failing it, but it is only in extreme (or wholly invented) situations that it is obvious that the methodological constraint is violated.

Take, for example, the conspiracy theorist. His case is easy, because he is delusional and clearly (and by stipulation) believes in silly things because such beliefs fit with his *other* silly beliefs. No amount of evidence will convince him that, e.g. fluoride is for dental health and not mind control. He ignores the actual data in favor of constructing layers of theoretical artifice to explain that the data itself is not representative of reality. While he certainly takes *himself* to have veridical beliefs, the conspiracy theorist's process of belief-formation is not, in fact, truth-seeking. If he were right, of course, his method would be laudable. If the explanatory connections that he identifies as explanatory really *were* explanatory, and not coincidence or what-have-you, then the conspiracy theorist would be vindicated. But since they are not, he is not, and we can conclude that his beliefs are not methodologically justified.

While we can conclude that a proper belief-formation method must be methodologically fallibilist as discussed above, methodological justification is not typically a property we can ascertain from the first-person perspective of rational inquirers. Methodological justification is essentially that constraint which ensures that a method of belief-formation really *is* truth-seeking. That is, in the long run and with repeated application, the produced belief network has true beliefs, so long as the environment is epistemically friendly.[6] Importantly, the criterion of methodological fallibilism ensures that local belief-formation methods are open to revision. Thus if method m_1 is found to produce false beliefs or otherwise fail at truth-seeking, it is replaced by method m_2. This should *not* be understood to imply that the beliefs formed by m_1

[6] It is part and parcel of the definition of "reality" that it is epistemically friendly, but such *a priori* arguments are not the only way to account for epistemic opacity in the world. This will be discussed in Chapter 5.

were not methodologically justified, despite their unreliable formation! The fact that m_1 was identified as faulty and revised indicates that there was a higher-order (i.e. global) belief-formation method at play. The higher-order method – the properly global method – really *is* truth-seeking and allows the produced beliefs to count as methodologically justified. When evaluating belief-formation methods for methodological justification, it is crucial that we take into account the criteria of method-selection by which they were chosen. If the (global) criteria for (local) method-selection select local belief-formation methods that seek the truth as best as they are able, the overall system is methodologically justified.

As an example, consider the following:

Potassium Si-anide: Sahar and Si. Sahar and Si are scientists in a chemical laboratory. They form beliefs by the exact same local method – they believe that any particular compound is composed of whatever the mass spectrometer says (it is a very advanced piece of machinery that requires no further analysis for the identification of most compounds). Unbeknownst to them, the mass spectrometer broke over the weekend. Although it is broken, the spectrometer luckily identifies a sample of potassium cyanide (KCN) as potassium cyanide. Sahar and Si accept this and believe that the sample is potassium cyanide, and this belief is coherent with their respective belief-networks. Are Sahar and Si methodologically justified in their belief that the sample is potassium cyanide?

The question as to whether Sahar and Si are justified in Potassium Si-anide cannot yet be answered. We know that they formed their beliefs that the sample is KCN based on the mass spectrometer's reading, and that the spectrometer is broken (i.e. the veracity of their beliefs is very lucky indeed). With only the information from Potassium Si-anide, we cannot determine if Sahar and Si are methodologically justified (and hence could know) that the sample is KCN. As such, we must expand the example:

Potassium Si-anide: Sahar and Si. Sahar and Si are scientists in a chemical laboratory. They form beliefs by the exact same local method – they believe that any particular compound is composed of whatever the mass spectrometer says (it is a very advanced piece of machinery that requires no further analysis for the identification of most compounds). Unbeknownst to them, the mass spectrometer broke over the weekend. Although it is broken, the spectrometer luckily

identifies a sample of potassium cyanide (KCN) as potassium cyanide. Sahar and Si accept this and believe that the sample is KCN, and this belief is coherent with their respective belief-networks. After analyzing a few more samples and receiving silly results (as one would expect from a broken mass spectrometer), Sahar and Si realize that the spectrometer is broken. Sahar begins to doubt whether the original sample was really potassium cyanide and verifies its composition as KCN by other means (but does not give this information to Si). Si simply maintains his belief that the sample is KCN. Are Sahar and Si methodologically justified in their beliefs that the sample is potassium cyanide?

With the further information available, it becomes apparent that there is a significant difference between the belief-formation processes of Sahar and Si. While Sahar correctly recognized that her belief that the sample was KCN was not reliably-formed, Si did not have this insight. Although both followed the same local belief-formation method of "Believe whatever the mass spectrometer says," Sahar correctly identified that method as unreliable and updated her belief in the relevant propositions accordingly. Si, although he admitted that the spectrometer was broken, did not so do. Sahar's super-method is methodologically fallibilist (and truth-seeking), while Si's is not. The truth-seeking nature of Sahar's super-method justifies those beliefs formed by locally non-truth-seeking means precisely because she is able to recognize the failings of her past local methods and integrate the relevant propositions into a belief-network that seeks truth. Sahar is justified in her belief that the sample is KCN, while Si is not.

Of course, we could further specify the conditions above such that Sahar follows some non-truth-seeking method (e.g. she believes dogmatically, perhaps) and Si follows an actually-truth-seeking method. We can weave an infinite number of alternate narratives around the toy examples to make them say whatever we like. Given the information presented, though, I find it intuitively plausible that Sahar is justified in her belief with respect to the composition of the sample, and that Si is not. I would go so far to say that insofar as Sahar *believes* that the sample is KCN, she *knows* that it is KCN. Her performing the analysis is an important *indicator* that her

method is properly truth-seeking, but it is not the double-checking that grants her methodological justification (perhaps through increasing the reliability of her local belief-formation method). That is, her methodological justification doesn't come from the reliability of her double-check method, but rather from the super-method she employed that allowed her to see the error of her past local method and revise it. In being able to embed her previous belief in an ongoing context of diachronic inquiry, Sahar shows that her global method revises itself (by whatever means seem the best to her at the time) in search of the truth.

Obviously, Potassium Si-anide is a Gettier example. It shows that while luck can mask the local results of a non-truth-seeking or unreliable method, it is not the *reliability* of the local method itself that is important to methodological justification, but rather our *response* to that luck and our evaluation of the method.

A recurring feature of the above analysis is the insistence on the truth-seeking capacity of the global method used. In cases where a local method is unreliable an agent can be methodologically justified if her higher-order method-selection procedure (her global method) allows her to improve upon her local belief-formation method by revising the relevant methodological commitments (i.e. if she can embed her belief in a genealogy of beliefs that explains how her current beliefs are a revision and improvement of her previous beliefs). This requirement is strikingly similar to my account of "true enough for 'true belief'" in Chapter 2.

In Chapter 2, I defined a belief as "true enough" for the attribution of true beliefs when the translation that belief into the language of the full descriptive truth is recognizable as the grounds for empirical success. While past theories are strictly-speaking false, the preservation of empirical predictive success (when it comes to scientific theories) and successful individual actions (when it comes to the truth of an agent's beliefs) will allow us to say what (if anything)

was *right* about previous beliefs. The relations that account for this positive evaluation of past beliefs can take many forms, and Barrett is correct to note that our account of what is "right" about a belief will be different for each of the theoretical virtues that turn out to be the *right* (that is, truth-conducive) theoretical virtues (2003, p. 12). The most obvious case is that of predictive success, as when we can recognize that "air" today means the mixture of gases around us, where "air" in the context of 17th-century chemistry was a single primitive element. Because our senses give it to us that "air" in both cases refers to that breathable gas around us however understood, we charitably understand that even though previous theories were woefully incorrect as to the nature of air, they at least identified something that *really exists*. In the case of caloric and phlogiston, we now understand that they are simply not a part of the world, although the properties and processes that they were posited to explain – heat and combustion respectively – are currently well-understood. "Air" is translatable across various theories as is "heat" and so 17th-century sentences containing these terms can count as true enough for true belief, and so "Air is breathable" and "water absorbs more heat than steel" count as "true enough." On the other hand, "caloric" and "phlogiston" find no home in the language of contemporary physics and so sentences like "Caloric exists" and "Phlogiston is an element" are false.

Just as the epistemically relevant understanding of truth in "true belief" requires a degree of charitable translation, so does the notion of methodological justification in "justified true belief." A method *m* is appropriately truth-seeking just in case we could make sense of its successes and failures from the point of view of the full descriptive truth. Some methods, like astrology, are not truth-seeking at all. We understand that astrology is not a reliable method for predicting the future. That said, it is plausible that any particular instance of astrological prediction could be embedded into an overall system of method-selection that makes it part of an

appropriately truth-seeking process. That is, if an astrologer's predictions are reasonably precise and come true every time, she is quite reasonable in continuing to employ the method. It is only if she lacks the capacity to revise her method to an appropriately truth-seeking process that we are able to call her belief-formation processes "unjustified." Thus if her astrological inquiry is so-embedded, we can say that those true beliefs it generated were appropriately methodologically justified. In the more likely case that the astrologer explains away incorrect predictions by appeal to bad psychic energy or what-have-you (or, even more likely, simply fails to make any precise predictions at all), her true coherent beliefs generated by an astrological inquiry are not justified precisely because they constitute the cessation of inquiry and the termination of the diachronic process of belief, action, doubt, and revision whose repetition constitutes a truth-seeking practice.

Without this fairly liberal understanding of justification, we rapidly discover that none of our beliefs are justified. As it turns out, it is extremely likely that none of our beliefs are true in the sense of the full descriptive truth. They are, of course, true-enough. Similarly, it is likely that none of our current belief-forming methods are perfectly reliable. After all, we *know* that our current best theories are false when taken together. Since we know that our best science is false, it begins to look like we are merely *lucky* when true (that is, true enough) beliefs are generated by appeal to science. More succinctly, we have good reason to believe that *all* our true beliefs are Gettiered.[7]

On the other hand, there *is* something right about our current best science, despite the fact that its sentences are false. We expect that many of its sentences, ontological commitments, and

---

[7] See Chapter 6 for a more thorough discussion of universal Gettierization

methodological commitments (e.g. parsimony, data priority, and so on) will be translatable into the language of the full descriptive truth in such a way as to identify some of these commitments as causes of local experimental success. Just as we must allow the ontological commitments of previous theories to vary when translating them into contemporary language, we must allow what we take to be appropriately truth-seeking methods to vary. In a phrase, our method must be *truth*-seeking at the higher-order (that is, our methods of belief-formation must be open to revision) and locally *true-enough*-seeking.

When we say that methodological justification is determined by the true-enough-seeking quality of a method and the criteria for its selection, we affirm that certain Gettier cases are genuine examples of methodological justification. In what follows, I will provide a rehabilitation of several popular Gettier cases (that is, cases of epistemic luck).

The Job: Smith and Jones have applied for a job. Smith has strong evidence for (a)"Jones will be hired" and (b)"Jones has 10 coins in his pocket." Smith accepts (c) "The man who will be hired has 10 coins in his pocket" on the basis of inference from (a) and (b). As it turns out, Smith is the man who will get the job and Smith also has 10 coins in his pocket. Is Smith methodologically justified in believing (c)? (Gettier, 1963, p. 122)

Clearly the method by which Smith forms his belief is not very reliable, as valid inference from false premises does not yield true conclusions with any particular regularity. To determine if Smith is justified, we must find out if his belief-formation can be embedded in a genealogy of beliefs that tends towards the truth. Once Smith gets the job and counts the coins in his pocket, we can be confident that he would be able to articulate his past beliefs in such a way as to account for the actual empirical success ("I took Jones to be the man who would get the job, and knew he had ten coins. But really I got the job *and* had happened to also have ten coins! How odd!") *and* the general unreliability of his method. If we assume that Smith is at all a reasonable person with respect to belief revision, then his case is easy to embed in the larger

truth-seeking practice that is inquiry, and he is methodologically justified. If, however, Smith believes in a way such that revision is not possible, he is not so justified.

Barcelona Brown: Smith has strong evidence for (a) "Jones owns a Ford." From this, he infers (b) "Jones owns a Ford or Brown is in Boston," (c) "Jones owns a Ford or Brown is in Barcelona," and (d) "Jones owns a Ford or Brown is in Brest-Litovsk." As it turns out, Jones does not own a Ford and Brown actually *is* in Barcelona. Smith has no idea as to Brown's whereabouts. Is Smith methodologically justified in believing (c)? (*Ibid.*)

Again, Smith's justification depends on how his beliefs can be embedded in ongoing inquiry. To the extent that from the point of view of the truth Smith could identify the successes and failures of his previous beliefs, he is methodologically justified. The fact that his beliefs were revisable in favor of *better* beliefs shows that his overall super-method is truth-seeking, despite his local unreliability. Of course, if Smith is unable or unwilling to revise his beliefs in the face of new evidence, he is not methodologically justified.

Barnsylvania: Henry is driving through the countryside. He sees a barn and judges that it is a barn. Unbeknownst to him, some high proportion of local barns are barn facsimiles. Although Henry identified a real barn, if he had set his gaze upon a fake barn, he would have judged it to be a real barn. Is Henry methodologically justified in believing that he saw a barn? (Goldman, 1976)

Once again, what we say about Henry depends enormously upon how he would react to learning that he is in Barnsylvania. If Henry is at all reasonable, learning about the high percentage of fake barns about would cause him to doubt his previous belief that he saw a barn. Either Henry would replace that belief with something like "I saw something that looked like a barn" or he would check to make sure that he really *did* see a barn. In either case, Henry is methodologically justified in his belief despite its local unreliability.

Sheepdog: Susan looks into a field and sees what she takes to be a sheep. She judges that there is a sheep in the field. What she took to be a sheep was in fact a rather ugly dog, but luckily there happened to be a sheep in the field hidden behind a stone wall. Is Susan methodologically justified in believing that there is a sheep in the field? (Chisholm, 1966)

Here, too, Susan turns out to be justified if she is at all a reasonable person. She says "There is a sheep in the field," but learning that she instead saw a dog would cause her to revise her beliefs. Learning about the dog and the actual sheep would cause even more revision of her beliefs. From the point of view of the truth, Susan would be able to say exactly how her belief was right (it was true, after all: there was a sheep in the field) and how it was wrong (it identified a dog as a sheep). This sort of misidentification parallels what we see when examining old scientific theories: phenomena are correctly predicted, but the predictions are not correct for the reasons that scientists think. Ptolemaic models of the solar system, for example, were extremely empirically successful, but not because the sun orbits the earth.

Calcuhater: Paul runs a computer program that adds two numbers. He has every reason to believe that it functions in a standard manner, calculating their sum and returning the result. He gives his inputs of 2970 and 3822, and the correct output (6792) is printed on the screen. Paul thus comes to believe that 2970 + 3822 = 6792. Unbeknownst to Paul, the program is actually a calcuhater that was programmed to output correct-looking but false results, but occasionally generates a correct result.[8] Is Paul methodologically justified in believing that 2970 + 3822 = 6792?

In keeping with the theme of this section, whether Paul is justified depends on other factors external to the example as stated. If the locally unreliable method of using a calcuhater to compute sums is embedded in a larger truth-seeking method, perhaps using computers to calculate sums when one's mental arithmetic is likely to err, then Paul's belief is methodologically justified and could be knowledge. If the super-method in which the locally unreliable method is embedded is *not* truth-seeking in this capacity – for example, if learning of the calcuhater's programming would not prompt Paul to revisit the belief in the absence of other

[8] To be more precise, a calcuhater only takes inputs between 1,000 and 10,000. It then generates a number between -5 and 5 and multiplies that number by ten. This number is the "hate." It then adds the hate to the correct sum and returns that value. We can expect that a calcuhater gives the correct result whenever the random number generator returns zero, about 9% of the time. See Appendix C.

confirming evidence – then Paul is not methodologically justified, and could not know. Insofar as Paul's belief is embedded in the context of inquiry, Paul is methodologically justified.

In discussing the above examples, it becomes readily apparent that methodological justification is as much a property of *believers* as of individual beliefs. Methodological justification is determined by the agent in question's approach to evidence and the generation of doubt when presented with evidence for their belief-formation method's lack of reliability. If a belief system (that is, the complex of belief, action, and intention that implies an agent whose outputs are best explained by the ABI model discussed in Chapter 2) does indeed approach the world in a way that embraces methodological fallibilism and is self-correcting with an aim towards the truth, we can call that belief-system a *knowledge-system* and we can call the implied agent a *rational inquirer*.

3. Entrenched beliefs

While methodological justification – being a rational inquirer– can generally be expected to be a fairly stable characteristic of a particular agent, it is reasonable that particular domains of knowledge – or even particular propositions – might be entrenched in an agent's belief network such that they are not vulnerable to doubt in the face of contrary evidence. Conspiracy theories, religious beliefs, and so on are paradigmatic cases of this, though there are certainly examples of scientists holding fast to outdated theories. In these cases, although one lacks methodological justification with respect to the proposition or network of propositions so tenaciously-held, this failing need not spread to their other beliefs. In this case, one would be a rational inquirer with respect to p (or subnetwork m), but not with respect to q (or subnetwork n). This section argues

that a belief's degree of methodological justification is determined by the relationship between that belief's vulnerability to doubt and the reliability of its method of formation by examining three cases of entrenched (i.e. not vulnerable to doubt) belief.

It is important to note that the original local method of belief-formation is not immediately relevant to whether or not one is justified in believing a proposition. Most of us learn an immense number of true propositions about the world well before we can be reasonably called rational inquirers, well before we develop the discerning eye that is part of inquiry. As children, most of us accept whatever we are told, socialized, or otherwise taught to believe without much (if any) thought. As we get older, we hone our ability to doubt propositions and to revise our beliefs in light of further evidence. We move from Peirce's method of authority to the scientific method, and can become rational inquirers (and the doubt that causes us to find a new method is evidence that we are indeed rational). But this does not mean that all the propositions learned in childhood are unjustified! Take a proposition that most of us learned in childhood: lit stoves are hot. The fact that this proposition was not methodologically justified when we began to believe it is quite beside the point. If we assume *now* that we are rational inquirers, then we have the ability to revise our beliefs (that is, to transform belief to doubt and inquire anew) with respect to lit stoves and their temperature. If we get good evidence that "Lit stoves are hot" is false, we can revise our beliefs. As such, we are justified in believing that lit stoves are hot insofar as lit stoves are indeed hot.

What then of my earlier comment regarding entrenched beliefs (or belief networks) that are not subject to doubt and revision? The attribution of methodological justification to a belief (and/or of rational-inquirer-status to an agent) is, like truth, a matter of retrospective charitable interpretation from the point of view of the full descriptive truth. If we were to have access to

the truth, we would be able to say precisely what local methods are reliable and what global methodological commitments are necessary for the global method to be truth-seeking. If all the data were in, we would know to what extent an individual's belief network is or is not amenable to doubt. Everyday experience teaches us that while many (or perhaps most) people are entrenched in some (or many, or most) of their beliefs, those beliefs are not completely invulnerable to doubt. Instead, those beliefs of which we are most sure are precisely those beliefs that lead us to think that the error is not in our beliefs, but rather in our sensory apparatuses and/or our evaluation of the data. And this, of course, is as it should be. We take all of our beliefs to be true, and we take our methods for forming beliefs to be those methods that *really do* yield true beliefs. Veridicality is a desirable property for beliefs to have precisely *because* true beliefs will not be contradicted by the state of the world – that is what is for them to be *true*, after all. Thus when we are certain about one of our beliefs, it is reasonable to be skeptical of contradictory evidence, particularly when the contradicted belief is supported by our *other* beliefs, which of course we also take to be true. To this end, the potential scenarios wherein an agent is *not* methodologically justified in the broad sense requires stretching the limits of credibility. For Paul to *not* be justified in Calcuhater, he would have to be such that he (1) Has no confirming evidence for his belief that 2970 + 3822 = 6792, but *also* refuses to reevaluate his belief that 2970 + 3822 = 6792 when he finds out that he has been operating a calcuhater. This case is extraordinary. While we can ascribe staunch failure to revise beliefs to various characters in our short examples, it is difficult to imagine a real, living human being behaving in such a manner.

Consider Alejandra, a professor of mathematics. Alejandra is absolutely sure that given Euclidean geometry the circle cannot be squared with a compass and straightedge. She is, of

course, correct. When a promising undergraduate student tells her that he has squared the circle, she is simply *right* to assume that her student (however promising he may be) has made a mistake. After all, von Lindemann proved the impossibility of squaring the circle when he proved that π is a transcendental number – Alejandra is safe in dismissing evidence that purports to contradict her belief that the circle cannot be squared with a compass and straightedge. Moreover, if she *herself* derives a theorem that implies the possibility of the squaring of the circle, she has good evidence via *modus tollens* that she made a mistake in her derivation procedure.

Alejandra shows us that there are two important senses in which a belief can be invulnerable to doubt. Alejandra's belief in the impossibility of squaring the circle with a compass and straightedge is of the first type, where the belief is invulnerable to doubt because it is *true* and could not easily (or, in this case, at all) be false. This category of entrenched beliefs encompasses logical truths, mathematical truths, nomological truths, and so on. There is, however, a second type of entrenched belief that we do not evaluate so positively.

To illustrate the second, less acceptable type of entrenched belief, consider Bikram. Bikram is a religiously devout genetics researcher. He is certain that he will be reincarnated after his death. He is also an impeccable scientist, and allows that any hypothesis in his laboratory might be disproved by future experimentation. In all non-religious aspects of his life, Bikram is methodologically justified by our current lights. When it comes to his religion, though, Bikram believes tenaciously instead of fallibly, and is not methodologically justified.

Of course, my statement that Bikram is not methodologically justified in his religious beliefs assumes that he is in fact *wrong* to believe in reincarnation. If Bikram is right, then my evaluation of him is quite faulty. Indeed, it might be the case that his method (perhaps

something like “Believe what the *Bhagavad Gita* says about reincarnation”) is more reliable than the methods he uses in his research (which range from “Believe what was written in my biology textbook” to “Believe the theory I helped to develop and which is the best explanation of the data generated by years of genetic research”). If the *Bhagavad Gita* really *is* a reliable source of information regarding reincarnation as Bikram takes it to be, then he is certainly methodologically justified in believing the relevant religious propositions. But let us assume that Bikram is wrong about this for the sake of argument.

Finally, consider Clint. Clint a conspiracy theorist, and he believes that the world is largely controlled by the British royal family, who Clint takes to be lizard-people in disguise. Any evidence that Alejandra or Bikram would consider to contradict Clint’s beliefs, Clint takes to be a further example of the lizard-people’s influence. When a scandal erupts in a political organization that Clint takes to be in collaboration with the British royal family, he confidently tells Bikram that this scandal merely proves that the lizard-people are smart enough to stage scandals to hide their true power. In short, Clint is the paradigmatic conspiracy theorist. To be clear, he is not schizophrenic or otherwise suffering from a treatable mental health issue; Clint is just silly.

Clint, of course, has many false beliefs. He is generally incredibly unreliable when it comes to interpreting the way that the world is and forming beliefs that are a good basis for action. He takes all his beliefs to be true, and interprets his frequent failure in action not as evidence that his beliefs are false, but rather that the British royal family really *does* run the world.

Like Bikram and Alejandra, Clint is not purposefully silly, nor does he see a reasonable alternative to his belief-forming processes. He forms beliefs according to what he takes to be the

most reliable way. Where Alejandra takes her knowledge that the circle cannot be squared to be sufficient grounds for dismissing her clever student's claims to the contrary, and Bikram takes his faith to be sufficient grounds for believing that he will be reincarnated, Clint takes his "knowledge" of conspiracy theories to be an appropriate basis for further belief-generation. Although Clint is obviously *wrong* about this by our lights, there is a deep structural similarity between his methodology and those of Alejandra and Bikram.

Alejandra, Bikram, and Clint all choose their belief-forming methods according to that which they take to be most conducive to the generation of true beliefs. That is, each of them uses methods that they *think* are reliable. In Alejandra's case, she discounts evidence based upon her belief in the veracity of mathematical proof. In Bikram's case, he takes certain passages of the *Bhagavad Gita* to be true and ignores that there is no direct evidence of reincarnation. In Clint's case, he takes his conspiracy theories to be true and a reasonable basis of belief. The difference among them is how we – roughly speaking, Anglophone philosophers and scientists – evaluate the actual reliability of their methods.

In the case of Alejandra, she is part of "our" community of inquirers. We share her fundamental beliefs with regards to the veracity of mathematical proof, and count her bases of belief as appropriate. Alejandra is justified by our lights, and she is a rational inquirer. Bikram is not much different. Although he is committed to particular religious tenets, nothing in the example gives us pause in attributing methodological justification to him. Because Bikram restricts his entrenched beliefs to the domain of religion, it is not too much to say that if he is reincarnated, then he was justified in believing that he would be reincarnated (even if the *Bhagavad Gita* is a work of fiction), and if he is not reincarnated then he was not so justified. The important tenets of inquiry – fallibilism and responsiveness to data – are intact in Bikram's

belief-forming methods in non-religious domains in a way that we can recognize, and so he is in general a rational inquirer. Clint, though, is a different story. He is outside of our community of inquiry, since we do not believe in his conspiracies, that the British royal family are lizards in disguise, and so on. To be sure, Clint *takes* himself to revise his beliefs in the face of contradictory evidence (perhaps he "learns" that the lizard-people can withstand temperatures as low as freezing when he reads a news article about the Queen going on a trip to Norway). By *our* lights, however, Clint's belief is generally not methodologically justified. His conspiracy theories are not a reliable method of belief-formation, and his certainty in his conspiracy-beliefs is unjustified. If he stumbles upon a true belief by this method – perhaps he correctly predicts a United States presidential election only on the basis of his conspiracies – we do not want to call this belief "knowledge."

The core of Clint's error and of Alejandra's success are the relationships between the certainty of their beliefs and the reliability of their belief-forming mechanisms. Alejandra is absolutely certain that the circle cannot be squared with a compass and straightedge, and her belief is based upon what we take to be a paradigm case of reliability: valid mathematical deduction. She is absolutely certain in her belief, but her belief-forming method is deductively valid, and so we can be certain that if she starts with true premises and does not make a mistake, she will end up with true conclusions. But the tenacity with which Clint clings to his beliefs does not match our evaluation of the reliability of his belief-forming mechanism. He is as certain of his beliefs as Alejandra, but the fixity of his beliefs outstrips the reliability of the processes by which they were formed. Indeed, the basis of applying the pejorative "conspiracy"

to Clint's theory of the world can be located precisely in the fact that his credence in his beliefs far outstrips the reliability of his belief-formation method.[9]

4. A new Measure of Justification

We are all methodologically justified in the broadest sense. When we *actually* doubt one of our beliefs, we revise it according to what we take to be the most reliable method we have for the generation of truths. Even in cases of extreme epistemic luck like The Job, Barcelona Brown, Sheepdog, Barnsylvania, and Calcuhater, the reading of the example in which the protagonist is generally reasonable yields that the protagonist is also justified in his or her beliefs. In a more narrow sense, however, we can evaluate the degree of methodological justification as the degree of proportionality between the reliability of their local belief-forming methods and the tenacity with which the proposition is believed. In more Peircian terms, there should be an inverse relation between the reliability of a particular local belief-forming method and the vulnerability of that belief to doubt. This section focuses on articulating a formal understanding of methodological justification.

If a belief is vulnerable to doubt, then it is revisable in light of new evidence. If a belief can be revised, it can be embedded into diachronic inquiry, no matter how silly it is. In measuring a belief's vulnerability to doubt, we also measure the degree to which that belief is amenable to inclusion in a truth-seeking practice and the degree to which we could expect that belief to be revised in the future. Of course, embedding a belief into diachronic inquiry does not

---

[9] This analysis matches the psychologists' definition of a conspiracy theory as "the unnecessary assumption of conspiracy when other explanations are more probable" (Aaronovitch 2009, p. 5; quoted in Brotherton, French & Pickering, 2013, p. 1).

mean that it is true (or that it was reliably-formed)! Rather, it means that the agent who so believes it can tell a story of the relevant belief: she can explain how she came to believe it and why she takes it to be the best basis for successful action. Days or hours later, she might change her mind. At this point, she would be able to describe what features of her previous belief are maintained and what features have been abandoned. In short, she could explain how her current system of beliefs accounts for the empirical successes of the now-jettisoned belief, and how it avoids at least some of the empirical failures of that belief. By pegging the reliability of an agent's belief-forming method to the tenacity with which that belief is believed, we ensure that we take our most reliably-formed beliefs most seriously. This, of course, says nothing about *actually* having reliably-formed beliefs, however. Instead, when beliefs are unreliably-formed, we are willing to give them up when they turn out to be false. This ensures that over time, our unreliably-formed false beliefs are replaced with true beliefs. As we develop our local belief-forming methods (and the methodological commitments that constitute our global truth-seeking practice) to more reliably produce truths, we recapture our old, unreliably-formed true beliefs and discover counterexamples to our false beliefs, ensuring diachronic movement towards the truth. Being methodologically justified means inquiring in such a way that one will, given enough time, be able to give a genealogy of one's system of beliefs from the silly notions of childhood through countless revisions and failed actions, finally ending with what one takes to be the truth.

It is rather difficult to come up with a sensible quantification of a belief's vulnerability to doubt. Vulnerability to doubt is easily defined from a given measure of tenacity or fixity, such that Vulnerability-to-doubt + Fixity = 1, but quantifying fixity is no easy task.

The most obvious candidate for a measure of belief fixity is degree-of-belief, often understood to be defined by what odds an agent would consider to be a fair bet (that is, a bet with an expected value of zero). This is the same measure as is used in subjective probability, Bayesian epistemology, and so on, and is much-discussed in the philosophy of disagreement. While I am amenable to this interpretation, it must be stressed that the fundamental concept in play in believing *p* to degree *d* is in behaving with tendency *d* to act as if *p* is the case. Betting is *one way* of behaving with tendency *d* that *p* is the case, but so are stating *p* with tendency *d* in appropriate settings, judging silently to oneself that *p* is the case with tendency *d*, and so on. Believing *p* with degree *d* is, given the nature of belief as a rule for action, the *same thing* as having the tendency to act as if *p* to degree *d*. The tendency-of-action interpretation does nothing to analyze degree of belief. I take it that this is because Eriksson and Hájek are right, and degree of belief is a primitive concept (2007).

Crucially, fixity (like betting behavior) should be understood as derivative from the generalized description of degree of belief as tendency to act as if a proposition were the case. When we believe a proposition with degree 1, we are certain that it is true. When we are certain – absolutely, unshakably certain – we reinterpret evidence in light of this certainty. Do note, of course, that English is a slippery enough language that we can construct apparent counterexamples quite easily. Most saliently, we do not balk at the sentence "I was certain that *x*, but I was wrong." The sentence, plausible though it may be, does not truly exhibit a degree of belief of one. When I understand myself to be wrong about *p*, I understand (roughly speaking) that it is more likely that ~*p* is the case than it is that all my evidence for ~*p* is mistaken. While it sounds reasonable to say "Sally was certain that Santa Claus existed before her mother told her the truth," the situation is better analyzed as Sally taking it to be more likely that Santa doesn't

exist than that her mother is lying to her in stating his nonexistence. Maintaining *p* in the face of evidence that ~*p* (or reinterpreting the evidence as supporting *p* and not ~*p*) is a way of believing *p* with a high degree-of-belief. Rejecting *p* at the first sign of trouble is a way of believing *p* with low degree-of-belief.

The above metric – the confirmation of an inverse relationship between reliability of method and the vulnerability of a belief to doubt – does not allow us to evaluate how justified we are in our *own* beliefs, or even the degree of justification of a particular agreed-upon belief within a community of inquirers – it is only useful in evaluating others' justification by our own lights (i.e. on the likely-false assumption that our *own* beliefs are both true and justified). We (i.e. Anglophone philosophers and scientists) might say that Bikram's belief in reincarnation, being based upon the *Bhagavad Gita* that we take to be highly unreliable, should be quite vulnerable to doubt. But Bikram's Hindu colleagues would probably disagree, even if they are not near as certain of their religious convictions as Bikram (or are even atheists raised in a predominately-Hindu culture). Even Clint takes his degree of belief to be proportional with the reliability of his belief-forming methods. After all, he forms beliefs according to his particular methodological commitments *precisely because* he takes those commitments to be the most conducive to forming true beliefs. The narrow sense of methodological justification is only useful as a comparison between two competing sets of methodological commitments, between two competing communities of inquirers (where the community in question might consist of only a single individual). The understanding articulated is itself an encoding of what I (and the community of inquirers of to which I belong) take to be a good method for obtaining true beliefs. If one had far different methodological commitments – for example, Descartes at the beginning of his Meditations would say that we should only believe that which is certain and that we should

believe it unshakably – then one would evaluate justification differently. Given that I take the method that I do in fact use to be the best method (otherwise I would get a new method), and that I take the beliefs that I do have to be true beliefs (otherwise I would change my mind), and given my thoroughgoing commitment to fallibilism, it is not surprising that the measure of justification I advocate is a measure of a fallibilist attitude.[10]

An easy formalization of the narrow sense of methodological justification described above is:

$$J_b = 1 - | R_{Mb} - F_b |$$

That is, the measure of justification of a belief is the absolute value of the reliability of an agent's belief forming method minus the fixity of that belief, the sum of which is subtracted from one.[11] The fixity of the belief is determined by the individual believer in question; the reliability of the method is determined by an external community of inquirers of which the believer is not a part. Justification by this metric is a positive number between 0 and 1, inclusive. A measure of one means that the agent has perfectly matched her degree of belief with the reliability of her method; a measure of zero means that the agent is perfectly certain of (i.e. perfectly unshakable in) a belief formed by a wholly anti-reliable belief-formation process. In stating the formula, I assume that all degrees of belief are greater than 0.50; otherwise, the particular belief *b* and its degree of belief *d* would be best described as ~*b* and 1 – *d* respectively.[12]

---

[10] One could also see this as a measure of open-mindedness. To the extent that one's belief is more fixed than the reliability of their method demands, one is closed-minded. To the extent that one's belief is more vulnerable to doubt than it should be, one is credulous.

[11] It seems implausible that the correlation would be as direct as this. I am perfectly willing to allow coefficients, the addition of constants, and so on. My point is about the *structure* of methodological justification – I do not think that I have hit upon the exact formula for its calculation.

[12] Of course, it might be useful to model degrees of belief less than or equal to 0.50 in some situations. For example, a rational person should assign a degree if belief of 0.05 to the proposition that a 20-sided die will roll any particular face. In modeling such a scenario, it might be useful to assign belief in each of the sides as receiving a credence of

Tying methodological justification to the J-measure, and not the flat reliability of the process being used, is extremely important. Using the J-measure allows processes that are unreliable by our current lights to have produced knowledge in the past (and accordingly allows us to have knowledge today, given the falsity of our best scientific theories). The history of science (and a moment's reflection regarding the similar structure of our own beliefs) gives us good reason to prefer this measure of justification. Consider once again early modern-era chemists. They took (broadly speaking) there to be four elements: air, water, earth, and fire. They believed in many silly (by our lights) things, the transmutation of (what we now know are) elemental metals among them. It would not be too much of a stretch to say that by contemporary standards the modern chemists were more or less *unreliable* in their predictions. But of course they could still have a high J-measure. To invent some numbers, let us say that the 17th century chemist's system of the elements allowed him to predict empirical events with 60% accuracy. If he recognized the successes and failures of his system, such that he believed the (untested) propositions of his system to a degree of roughly 0.60, then his J-measure is relatively high. This is particularly important when we consider that the conceptual schema by which we approach the world counts as part of the method by which we interrogate it. At least *some* of the error in our modern-era chemist's practice is due to the word "air" on his lips simply meaning something *different* than it does today. The measure of reliability must not be understood as "Delivers the full descriptive truth 60% of the time," but rather "Delivers 60% of the truth, parceled in various

---

.05, as opposed to assigning the belief that the die will *not* roll a particular side a value of .95. Such transformations are not inherently problematic, despite the fact that it sounds strange to count a degree of belief of 0.50 as *belief* (as opposed to doubt).

ways."[13] For example, when the modern chemist says "2 + 2 = 4," he means something very close to what we mean today.[14] When he says "Air is transparent," he means something that is rather different from what we mean when we make the same utterance. Much of his unreliability is not due to his methods being empirical *failures* in the sense that he expected *p* and got ~*p*, but in the fact that there are fundamental flaws in his understanding of the world. If he expects combustion from an experiment and the experiment indeed yields combustion, the chemist's interpretation of this event as e.g. confirming the presence of phlogiston indicates the unreliability of his method.[15] It is intuitively implausible that "Johann Becher knew that air is transparent" is false. Measuring the appropriate fixity of belief instead of the reliability of that belief's formation-method allows us to articulate the sense in which past inquirers were justified in their beliefs: the fixity of their beliefs did not outstrip (or fall behind) the reliability of their belief-forming methods.

The J-formula has a further implication: those beliefs which are held to a sufficient degree to count as candidates for knowledge (perhaps those whose degree-of-belief is greater than 0.50) are analytically reliable when the J-measure is high. If we fix the J-measure at one, then the only possible knowledge-candidate beliefs are those that are also reliably-formed.

Although the J-formula for justification is useful, one might take an underestimation of the reliability of one's own belief-forming mechanisms to be epistemically better than an over-

---

[13] This will become vital in Chapter 6, where I develop a measure of knowledge or approximation thereof. Because the reliability of a method is determined by the closeness to the truth of the sentences it generates, we can expect that with access to the truth – at the end of inquiry – our belief-forming methods would be perfectly reliable.
[14] Though of course the system that he uses allows for the truth or negation of the theorems that have been proven in the interceding four centuries.
[15] Though he *thinks* that his method is quite reliable. A practical moral from this discussion might be that we should try to be more fallible than our estimation of our own reliability indicates, since vagueness and conceptual error that we are incapable of recognizing will probably bring down the measure of our reliability from the point of view of the full descriptive truth. In this sense, knowledge favors the cautious.

estimation. That is, it might be better to be overly-cautious than credulous. While *prima facie* plausible, such an understanding of justification would be incorrect. Justification measures the flat difference between the reliability of an agent's belief-forming method and her evaluation of that method as reflected in the degree to which she believes the relevant proposition. We can expect that rational people assign credence of 0.50 to either side of a fair coin coming up on a coin flip; one who assigns a credence of 0.48 to each side is exactly as incorrect as someone who assigns a value of 0.52 to each side. Indeed, assigning 0.48 to one side just *is* assigning 0.52 to the other – that is how fair flips of a coin work. It is extremely hard to see how overcautiousness – perhaps assigning a value of less than one to known-valid deductions from certain premises – is superior to its opposite.[16]

A further objection might come up: the J-formula seems to simply have wholly counterintuitive results. According to it, Ralph would be fully justified in assigning a credence of .50 to "2+2=4" if the method he uses is "Flip a fair coin and believe the proposition if heads shows." This seems flatly wrong!

I contend that the above example seems incorrect only because the obviousness of the believed proposition outstrips the limitations of the silly method described. That is, if Ralph *really* were to believe that 2 + 2 = 4 *only* on the basis of the coin flip, then it is not clear why he should believe it with credence of greater than .50. Of course, it is a fact of being a human that we accept propositions that accord with the rest of our beliefs. It is reasonable to hold that coherence with our other beliefs accords some weight to true propositions if we assume that we

---

[16] Though the presence of vagueness and error in our current conceptual schema does mean that we should give less credence to our beliefs than our evaluation of their formation indicates, this does not imply that it is *ceteris paribus* better to estimate too high than too low relative to the actual reliability measure. It merely says that we tend to overestimate the reliability of our current methods such that aiming low is a good strategy if we want to maximize the J-measure of our beliefs.

have mostly true beliefs. But *that* is not the method described in the example. In the example, Ralph believes *solely* on the basis of a coin flip, not on the basis of a coin flip *and* the ability to form beliefs that are coherent with our other beliefs.

It is also extremely *obvious* to us that 2 + 2 = 4. But if this obviousness is to count as part of a counterexample, then Ralph must have access to it as well. The method so-described – believing obvious mathematical propositions solely on the basis of coin flips – is nearly nonsensical. If we are to understand Ralph as anything close to a human being, he would be much better described as believing 2 + 2 = 4 on the basis of its coherence with the rest of his beliefs (whatever *they* are), with his understanding of the words "two" and "four" and his ability to count; the coin flip is rather irrelevant to anything we can recognize as a live possibility for a belief-forming method. For obvious necessary truths their obviousness and necessity swamps any stipulated belief-forming methods in thought experiments to such a degree as to render the examples counterintuitive. That is, we cannot put ourselves into Ralph's shoes, because we simply cannot imagine what it would be like to believe "2 + 2 = 4" solely on the basis of a coin flip. While there is no obvious *reason* that such a strange character as Ralph would not be maximally justified in such a situation, the result seems unintuitive to us precisely because the situation itself is incredibly strange.

5. The J-Measure Tested

Consider again our cases from Section 2: Potassium Si-anide, Calcuhater, Barnsylvania, Sheepdog, and Barcelona Brown. In Potassium Si-anide, Si incorrectly takes his belief in the mass spectrometer's reading to be reliable. While Sahar initially does the same thing, she

demonstrates the proportionality between the fixity of her belief and the reliability of the method by which it was formed by, in fact, re-evaluating her belief in the sample's composition. By reading the example in light of the justification measure above, it becomes obvious that while Si has *some* level of justification - he is not, let us assume, completely certain of his belief and immune to *all* doubt – his overestimation of the reliability of the (broken) mass spectrometer's output is epistemically problematic. This section applies the J-measure to the remaining examples, showing that it yields intuitive results.

Similar to Potassium Si-anide, the most obvious reading of Calcuhater has Paul assigning an inappropriate level of credence to his belief that 2970 + 3822 = 6792. The method by which Paul forms this belief – using a calcuhater – is about 9% likely to produce true beliefs, but Paul takes himself to be using an extremely reliable method and accordingly believes that 2970 + 3822 = 6792 quite tenaciously. As it happens, 2970 + 3822 really *is* 6792, but Paul's method is not to thank for this fact. If Paul were to be extremely cautious in assigning probabilities (perhaps he knows that there is a prankster about that loves replacing calculators with calcuhaters), he would be more justified in his belief.

In both Barnsylvania and Sheepdog, Henry and Susan's justification (or lack thereof) heavily depends upon how resolute they are in their beliefs that they are seeing a barn and a sheep respectively. To the degree that Susan (implicitly or otherwise) understands that she might mistake various quadrupeds for a sheep and is open to revising her belief, she is justified in believing that there is a sheep in the field. To the extent that Henry's judgment that he sees a barn is vulnerable to doubt and can be embedded in ongoing inquiry, he is justified. To the extent that either of them over- or underestimate the reliability of their sensory apparatus and thus forestall ongoing inquiry, their degree of justification is reduced.

Barcelona Brown presents an interesting and complex case. The degree of belief that Smith should assign to "Jones owns a Ford or Brown is in Barcelona" should only be marginally higher than the degree of belief that he assigns to "Jones owns a Ford." Smith's local method is, in actuality, random guessing, but he takes it to be deduction from true premises. As such, his degree of justification for "Jones owns a Ford or Brown is in Barcelona" is rather low.

Here, as is so often the case, the purported familiarity of the scenario described in Barcelona Brown leads us to import assumptions not specified by the case itself. As described in the preceding paragraph, it looks like Smith is not justified in his belief. However, he is not a total buffoon: he has lots of good evidence for his belief that Jones owns a Ford, but evidence as to Jones's deceit would certainly give Smith cause to re-evaluate his beliefs. In discussing the case, it is easy to gloss Smith's method as extremely unreliable (and so the appropriate fixity of the belief should be low, perhaps 0.51 comes to mind) and his evaluation of it as near-certain (perhaps something like 0.98). But, of course, he is not so tenacious in his belief that Jones is a Ford owner as to be completely silly, for the make of someone's car is an easy thing to get wrong. Given that Smith's degree of belief for the disjunction should be the sum of his degree of belief in the two (independent) disjuncts, and that he has no belief in in "Brown is in Barcelona" whatsoever, it begins to look like the upper limit for his credence is not as high as we might have initially thought – its maximum is his credence in "Jones owns a Ford." If this is the case, the final measure of Smith's justification begins to creep up, and he is more justified than we thought. To the extent that Smith is reasonable, and fallibilist, and able to revise his theories on the basis of new information and appropriately change his beliefs and account for the empirical failure of his previous beliefs – in short, to the extent that Smith's belief in the disjunction is embedded in diachronic inquiry – he appears to be justified. He is not, of course, *as* justified as

he would be had strange conditions not obtained, but this is precisely the benefit of using a non-binary understanding of justification: we can make distinctions between various cases without having to hold that e.g. luck is rare and mysterious and knowledge-destroying, as opposed to a feature of every belief and item of knowledge conceived throughout the history of our species' inquiry.

Barcelona Brown has a further complication: it is not clear that Smith believes "Jones owns a Ford or Brown is in Barcelona" in the first place. Part of the example is that Smith does *not* believe "Brown is in Barcelona"– he chooses it at random as part of a logic game. Beliefs are rules that we take agential systems to follow – they are postulates that we make as an explanatory aid in understanding what we take to be systems that are best described by an ABI model (recall the discussion in Chapter 2). For "Jones owns a Ford or Brown is in Barcelona" to be a good action-rule to posit for the system that is Smith, it seems as if we should have some reason for positing it as opposed to the simpler "Jones owns a Ford." The extra disjunct "or Brown is in Barcelona" does not give the rule in question any more power or predictive capacity, and as such interpreting Smith's actions to follow a rule that is something like "Behave as if Jones owns a Ford or Brown is in Barcelona" is a rather poor interpretation of the system. Much better would be to understand Smith as believing that Jones owns a Ford and having no particular belief with respect to Brown's location (as indeed is the case *ex hypothesi*) instead of adding in admittedly useless disjuncts. The case is rather similar to Peirce's complaint regarding whether we say that a stone on the bottom of the ocean (and destined to stay there) is brilliant or dull. "The whole conception of [a] quality…," after all, "lies in its conceived effects" (1878, p. 33); given that Barcelona Brown explicitly states that Smith has no belief one way or the other with regard to Brown's location, there can be no effect of positing the disjunctions as a belief to Smith

over the atomic sentence. Smith isn't justified in believing the relevant disjunction, but this isn't due to any problem in assigning degree of belief or using an unreliable method. He isn't justified in that belief because he *doesn't believe the proposition at all*. Beliefs are theoretical posits, and the disjunction-belief is not a good explanation for Smith's actions. The Gettier problem that the example is usually taken to identify is quite beside the point.

In each of the cases discussed, I have been careful to say that the various characters are justified to some extent or another. In giving a degree interpretation to justification, we affirm that justification (like most things) is not a binary affair. It makes little sense to say that Henry *is* or *isn't* justified, because justification is a matter of degree. This is not so surprising, and I think that it is in accordance with our general intuitions about language and the world that justification is a matter of degree and not all-or-nothing. While reliability is a component of justification (see e.g. Sosa, 2009; Sosa, 2011; Goldman, 1976; etc.), it is not all there is to it. This is precisely because belief itself is not all-or-nothing: we hold more tenaciously to some of our beliefs than others. Methodological justification comes in correctly matching the tenacity with which our beliefs are held with the tenacity with which the world forces that belief upon us.

6. Implications on Disagreement

The J-measure has rather interesting consequences in the epistemology of disagreement. There are two primary stances regarding the rational thing to do when faced with disagreement from an epistemic peer (that is, one who has the same evidence and similar cognitive abilities as ourselves). On the one hand is the Equal Weight View (EWV), in which we give our disagreeing disputant's testimony equal weight with our own, and update our degrees of belief

accordingly (Christensen, 2007; Elga, 2007). In effect, we will usually "split the difference" between our two views. If we are on precisely opposite sides of a binary proposition (i.e. I attach credence 0.75 to *p* and my disputant believes ~*p* with the same credence) this amounts to total uncertainty. This approach can be generalized to non-peers (that is, those who are more or less reliable than we are in generating true beliefs relative to a particular domain) as the Proportional Weight View (PWV) (Elga, 2007). According to the PWV, we should weight the testimony of disputants in proportion to their epistemic credentials. This view is intuitively plausible, as there appears to be no good reason to assign different degrees of belief to the testimony of epistemic and evidential equals, even in the case that I am one of those individuals (Christensen, 2007; Elga, 2007). This leads to the uncomfortable situation, argued against quite forcefully by Simpson (2013), wherein important beliefs are to be constantly reevaluated merely on the basis that someone disagrees with us.

The opposite position, wherein one maintains their degree of belief and simply ignores the testimony of epistemic peers, might be called the Steadfast View (SV) (Kelly, 2005). Although it is less common in the literature, SV has the benefit that it allows for continued public discourse (which one might think is a social good). As Mill says, after all,

> As it is useful that while mankind are imperfect there should be different opinions, so is it that there should be different experiments of living; that free scope should be given to varieties of character, short of injury to others; and that the worth of different modes of life should be proved practically, when any one thinks fit to try them. It is desirable, in short, that in things which do not primarily concern others, individuality should assert itself. (2006, p. 65)

To do justice to the discussion, it is first useful to define *peer-hood*. As the question is what one should do when an epistemic peer disagrees, we must know what is meant by "peer."

Kelly (2005), Christiansen (2007), Simpson (2013), and de Ridder (2013) follow Gutting (1982) in roughly understanding peers as having access to identical evidence and possessing similar cognitive abilities and epistemic virtues. Elga's view is somewhat more nuanced: I consider myself an epistemic peer with someone with respect to a particular question only if I expect that if we disagree, we are both equally likely to be correct. The received (though modified) view from Gutting sketches Elga's understanding in broad strokes, but Elga's definition is clearly superior in that it offers a precise definition over an appeal to character traits while capturing the very same intuitions about peer-hood.

If, then, I take my friend to be equally likely to be wrong as myself when we disagree, what should I do when such a disagreement comes up? For a clear case, consider Elga's example of a horse race: My friend and I have equally-keen vision and are attending to the race equally well, and the race ends in a photo finish between Alphonso and Beauregard. I believe that Alphonso won the race, and my friend believes that Beauregard won. Let us say that we each believe the relevant propositions to a reasonably high degree. In this case, it makes little sense to privilege my belief over that of my friend or *vice-versa*. In such situations, I should revise my degree of belief that Alphonso won down (and my friend should similarly revise her belief that Beauregard won). Given two equally-reliable sources of information, neither should be privileged and my degree of uncertainty should rise.

The type of "clean" disagreement mentioned above can hardly be called disagreement at all. "Disagreement" has connotations of consideration, evaluation of arguments and evidence, and so on. The horse-race example isn't an *easy* example of disagreement, but something else entirely. It is an example of contradictory data, to be sure. To sharpen this point, consider the rational response if I am not a party to the horse race, and instead two friends (with equally-keen

vision, paying equal attention to the race, and so on) give me conflicting reports about the winner of the race. I can be quite sure that *either* Alphonso or Beauregard won, but I have no reason to privilege one outcome over the other. This is entirely disanalogous with genuine cases of disagreement, wherein I disagree with a disputant regarding the existence of a supernatural being, or the permissibility of abortion, or something similar. Lowering my credence in a merely perceptual belief – as when my friend tells me that Beauregard crossed the finish line first – is simply *not the same thing* as lowering my credence in a reasoned conclusion based upon my best evaluation of the available evidence because of disagreement.

It is here that Elga's definition of peer-hood can be marshalled in Kelly's favor. Kelly's point is that disagreement about *substantive matters* (even among peers) is not, by itself, a good reason for me to lower my belief in my reasoned position. Kelly's point is that it is precisely *this sort of question* and my interlocutor's response to it that *makes* him or her a peer. But a "peer" by Kelly's lights is someone who is equally intelligent and has the same evidence that I do. And *this* does not mean that she is as likely as I am to be right in certain domains. As Elga notes, in cases of "insane answers," we are given reason to revisit the circumstances of evaluation, and potentially conclude that our disputant is not, indeed, as likely as we are to be correct. If we assume that religions *A* and *B* are mutually contradictory, and that proponents of each are equally intelligent, cognitively able, and so on and each have access to the very same evidence, it is quite reasonable for *A* to consider *B* an epistemic inferior with respect to religion ("She believes in reincarnation! That's crazy!") and for *B* to similarly conclude that *A* is epistemically inferior for the very same reason ("He doesn't even believe in reincarnation! That's insane!").

While Elga and Kelly appear to be in direct contradiction, we must remember that PWV *only* applies in cases that *really are* symmetrical. In the case of simple perceptual disagreement,

it appears that the case for Alphonso is precisely the same as the case for Beauregard. But in substantive– political, philosophical, religious, and so on – questions that are embedded in networks of related propositions that are also in dispute, PWV simply does not apply. In *these* cases, Kelly's moral is applicable: A social democrat need not worry solely on the basis that a libertarian says "Taxes are theft," and a religious person need not re-evaluate his belief in the supernatural because an atheist disagrees. The PWV can only apply in such cases by following the path of inference to its root, to the cluster of disputed propositions that entail the substantial disagreement about a particular case. Once *this* is reached, it might be possible to consider a disputant an epistemic peer in Elga's sense that she is equally likely as us to be correct. But once we have performed *that* step, the success (by our lights) of the theory in dispute weighs in favor of keeping us steadfast in our belief. The argument for accepting the axiom of choice is not that it is *true*, but that the mathematical structure it allows is richer, more useful, and indeed more beautiful than the structure that does not include it. Similarly, the social democrat's view that (say) justice is fairness yields what she takes to be *true* propositions about labor, the proper structure of society, and so on. Since her axiom yields true (by her lights) results, she does well to accept it. And the minarchist libertarian is in precisely the same position: what *he* takes to be the set of true propositions follows from his acceptance of *his* axiom (perhaps "justice consists of Nozickean justice in holding"), and so he does well to maintain his belief in it. That the social democrat rejects the minarchist libertarian's propositions is just to say that she does not expect that minarchist-libertarian systems are at all likely to produce normative judgments that are true, and so she cannot regard the minarchist libertarian as her peer. But of course this cuts both ways: by the minarchist libertarian's lights, asking the social democrat is a completely unreliable (and perhaps anti-reliable) way to form beliefs about social justice. The two are simply not

epistemic peers with respect to politics in Elga's sense, although they are equally intelligent, approach the argument with equal integrity, are equally sincere, and so on and so count as peers in the Gutting-derived sense.

I have made a short intuitive argument for why we should adopt the PWV in clear cases of peer disagreement like perceptual belief, in which the reliable formation of a single proposition (or cluster thereof) is at issue. We should remain steadfast with respect to substantial questions that involve large networks of interrelated propositions, but this is because the import of such questions render substantial peer disagreement extremely difficult. The fuzzy distinction between large and small propositional networks can be understood in Peircian terms as belonging to the same community of inquirers. Social democrats and minarchist libertarians approach the world in fundamentally different ways, and have substantially different views regarding justice, human flourishing, and so on. Each community works according to what it takes to be the most reliable local belief-formation methods, and revises its local methods in accordance to what it takes to be the best global truth-seeking method. Insofar as there is a fact of the matter with respect to justice, human flourishing, and so on, we can expect that one of these positions is closer to the truth than the other. Libertarian minarchists can reasonably disagree with each other regarding local matters of what is best for human beings, and can engage other minarchist libertarians in a mutual search for the truth (the truth, of course, that they expect will be some recognizable variation of minarchist libertarianism). The exact same can be said for social democrats: their commitment to social democracy is precisely their expectation that a rough version of social democracy will be recognizable in the full descriptive truth about justice, human flourishing, and so on. But insofar as social democrats expect that

minarchist libertarians are *wrong*, there is no relevant sense in which they expect minarchist libertarians are as likely as themselves to be correct when a disagreement arises (and *vice versa*).

Application of the J-measure gives us precisely the results that we would want in such cases. The J-measure is defined as one minus the difference between the reliability of a belief-formation method and the degree to which we believe that proposition. If social democratic theories turn out to be rather reliable, and we are rather confident in them, then we are methodologically justified in believing them (and similarly for minarchist libertarian theories). Presumably, most individuals have fairly strong beliefs regarding which of the two positions is closer to the truth, is supported by current empirical evidence of human flourishing, and so on, and presumably one of these theories is indeed closer to the truth. Whichever theory of justice is correct, it seems that the alternate theory is held despite the actual evidence to the contrary. Whichever is correct, the *in*correct theory appears to be believed not in accordance with data, but dogmatically (or tenaciously, or according to an *a priori* method). The incorrect theory seems to generally be believed to a high degree, while its reliability is quite low. Whichever theory is incorrect, its propositions are not believed in a methodologically justified way precisely to the degree that they are tenaciously believed and unable to be embedded in the context of diachronic inquiry. Whichever theory is correct, its adherents would do well to increase the J-measure of their beliefs. Thus they would be more likely to hold on to those beliefs they possess that are true, and reject those that are false.

The J-measure does fairly well, then, in cases of substantive and pervasive disagreement. It is similarly adept at dealing with cases that can be covered by PWV. To return to the horse case, if neither my friend nor I have any prior beliefs about which horse would win the race, and we both get at 90% of the truth with respect to judging horse races, and we both initially believe

that our chosen horse won the race with a credence of 0.90, we would both do well to revise our beliefs downward to withholding judgment. And this is analytic given Elga's definition of epistemic peer-hood: my friend is my peer with respect to the horse race if and only if the probability of Alphonso winning, *given* our disagreement, is 0.50. We disagree. So the resulting probability must be 0.50. By the very definition of peer-hood, it is indisputable that the reliability of my perceptual faculties given disagreement is 0.50. Thus if I am to maximize the J-measure of my beliefs, I should believe that Alphonso won to degree 0.50.

One might wonder, though, what the J-measure says in cases of Gutting-esque peer-hood, where we do *not* expect that our disputant is equally as reliable as we are, but that he is just as cognitively able, intellectually honest, operates from the same evidence and so on. In *these* cases, it is not clear at all that epistemic peer-hood is a particularly relevant feature of the world. In this sense, nearly everyone is our peer. Those who approach the world in significantly different ways – that is, those who are not in our community of inquiry with respect to particular topic or domain – might be as intellectually virtuous and intelligent as we want, but they get things substantively *wrong* by our lights. Just as it would be silly for me to substantially revise my system of beliefs in the face of a computer program that unreliably or anti-reliably spits out propositions about the future, it would be silly for me to substantially revise my belief based on the mere fact of disagreement by Gutting-peers. This, of course, leaves open careful and honest consideration of actual *arguments* given my epistemic peers external to my community of inquiry. If I am indeed as intellectually virtuous as stated in the examples, I will carefully consider any arguments they present with respect to the axioms that structure a particular domain (e.g. "Justice is fairness") and the particular judgments that result from such axioms. Meaningful discourse is possible across communities of inquiry – it is just that the mere fact of disagreement

does not require downward revision of my credences. Thus in the case of Gutting-peers, I side with Kelly: disagreement of Gutting-peers is not epistemically significant precisely *because* one can be a Gutting-peer while being unreliable (or anti-reliable) in a particular domain. Peer-hood in the more precise sense first articulated by Elga, though, *is* epistemically significant. The Proportional Weight View is the right rule to follow in such instances. In both cases,

7. Conclusion

Methodological justification is the degree to which a belief can be embedded in an ongoing truth-seeking practice. The hallmark of such a practice is its fallibilism with respect to its own methodological and ontological commitments, as for a method to be properly truth- (as opposed to fiction-) seeking, it must be willing to alter itself to better fit the data that constitutes reality. This understanding of methodological justification, formally expressed as $J_b = 1 - | R_{Mb} - F_b |$, reveals that justification is best considered as a scalar measure such that while characters in various thought-experiments are justified to a greater or lesser degree, a complete lack of methodological justification is extraordinary. By our lights, an interlocutor that belongs to a competing community of inquiry has a low degree of methodological justification (and by her lights, we ourselves are not well-justified). Insofar as competing communities explain the very same data or the veridicality of their competing claims are in dispute, the existence of epistemically virtuous agents with opinions different from our own does not constitute a threat to our own degree of justification (when considered by our own lights). To the degree that I can agree with my interlocutor on theoretical virtues, observational data, and so on – that is, to the degree that she belongs to my community of inquiry – I must take her disagreement seriously

and lower both my credence in the disputed belief and my assessment of my own reliability if I am to maintain a high degree of methodological justification.

CHAPTER 5: RELIABILITY AND THE J-MEASURE

In the previous two chapters, I argued that there are two types of justification, synchronic and diachronic. The first is best described as *coherence*. Coherence is the measure of how a system of beliefs "hangs together," and our current best cognitive science suggests that humans do, in fact, believe those propositions that best cohere with our previous beliefs. The second type of justification is *methodological justification.* A belief is methodologically justified to the extent that it is embeddable in the ongoing truth-seeking practice of inquiry. For inquiry to be truth-seeking, it must be fallibilist: it must take its own methodological and doxastic commitments to be revisable in the face of unexpected results. Of course, inquiry should not be *too* fallibilist, lest we abandon our true beliefs in favor of convenient fictions. A measure of the appropriate "embeddability" of a belief in a diachronic truth-seeking practice – that is, its measure of methodological justification – is

$$J_b = 1 - | R_{Mb} - F_b |$$

where R_{Mb} is the reliability of the method of belief-formation and F_b is the firmness or fixity (resistance to doubt) of that belief in a doxastic network. Fixity can be easily understood as degree of belief, since the more certain we are of a belief the less likely we are to reject it, and the more likely we are to interpret novel data in light of its (presumed) truth. On this understanding, a belief is maximally justified just in case our degree of belief in it (that is, the firmness of that belief) is identical to the reliability of the method by which it was formed.

While the reliability of a belief-formation process is regularly understood to be a component of justification and/or knowledge (see, for example, Sosa, 1996 and 2000; Greco, 2000; Leplin, 2007), the *generality problem* – the problem that it is rarely clear precisely what

belief-formation method was used in any particular case – has proved to be a difficult challenge to reliabilist theories (see, for example, Baumann, 2009; Christiansen, 2007; Alston, 1995; Feldman, 1985). This section seeks to solve that problem and its more general case that constitutes the *reference-class problem* as they apply to theories of justification and knowledge which rely on reliability. Call such theories *reliabilist,* collectively *reliabilism*.

I will first argue that modal understandings of reliability fail, and that we are better off going with a probabilistic notion of reliability on pragmatic grounds. I will then attempt to answer the reference-class problem (of which the generality problem is a special case) by (1) Noting that the *actual* method we use in forming beliefs is well-defined and (2) Noting that we have a large store of beliefs that were generated and validated by that very process that constitutes an appropriate reference-class.

1. Modal Reliability

Sosa (2009, 2011) gives a generally reliabilist account of knowledge, though he is quick to note that it is better to speak of belief in terms of epistemic virtues than reliable processes so as to have a ready answer to the value problem. This aside, he states that "nothing is a cognitive virtue unless it is a truth-conducive disposition," (2011, p. 135). While he is not explicit, Sosa seems to rely on a modal notion of reliability, such that a belief-formation process is reliable just in case it produces true beliefs in some proportion of close possible worlds.[1] This is also the approach that Goldman takes (1976, 1986, 1988), and is similar to the covariantist notion of justification championed by Nozick (1981), whose finer points were debated in the dispute

[1] Indeed, see Sosa's (1996) and (2000) for explicitly modal understandings of reliability in terms of safety.

between safety and sensitivity (see, for example Comesaña 2005, Baumann 2008, Greco 2012). However it is worked out, modal reliability necessitates a class of possibilities – possible worlds – ranked according to a closeness metric. Since there is no clear way to differentiate the closeness of possible worlds without begging the question for or against a particular scenario as "close enough" to merit consideration, modal reliability is a non-starter.

To sharpen the point against modal reliability, consider the generality problem (Feldman, 1985). The generality problem is, essentially, that any particular belief-formation method one uses is a token of many different types. Recall the familiar case of Barcelona Brown, modified from Gettier's (1963):

Barcelona Brown: Smith has strong evidence for (a) "Jones owns a Ford." From this, he infers (b) "Jones owns a Ford or Brown is in Boston," (c) "Jones owns a Ford or Brown is in Barcelona," and (d) "Jones owns a Ford or Brown is in Brest-Litovsk." As it turns out, Jones does not own a Ford and Brown actually *is* in Barcelona. Smith has no idea as to Brown's whereabouts. Is the method by which Smith formed his belief that (c) reliable?

Is Smith's belief-formation method above best described as "inference," "guessing," "valid inference from false premises," or something else? Valid inference is intuitively a reliable procedure, but guessing is not. So does Smith have a reliably-formed true belief that Jones owns a Ford or Brown is in Barcelona?

The answer depends on whether or not there is a single, determinate method that Smith uses. If there is no single correct way to describe Smith's belief-formation process uses, then there is no fact of the matter with respect to his reliability. And if *this* is the case, then there is no fact of the matter with respect to any measure of justification that relies upon reliability. Such a result would be most unwelcome to many epistemologists.

A more general version of this problem – the reference-class problem – also applies. The reference-class problem is the difficulty of determining the class from which a particular sample

is drawn. For example, if I want to know the likelihood of a particular car suffering catastrophic failure, do I consider the frequency of catastrophic failure in all cars, cars of a particular make and model, cars of a particular year, or something else? A modal account of Smith's reliability would ask something like "Is the proportion of true beliefs to false beliefs generated by Smith's method (whatever it might be) greater than some ratio *r*?" To determine this, there must be some set of interesting worlds. If the proportion of true to false beliefs that Smith has in these worlds is greater than or equal to *r*, he is reliable. Otherwise, he is not. But which worlds shall we use? There are an infinite number of possible worlds (one might think) that differ in all sorts of ways wherein Smith indeed believes "Jones owns a Ford or Brown is in Barcelona" and Brown is indeed in Barcelona. Are those worlds closer or further away than the worlds where Smith guesses Belgrade instead of Barcelona? Are they closer or further than those worlds where Smith performs the exact same inference, Brown is in Barcelona, and Jones actually does own a Ford? How about the worlds where Jones owns a Ford and Brown *isn't* in Barcelona? The ordering of worlds – and what counts as "close enough" for a world to merit consideration as part of the overall reference class – entirely determines whether or not a particular method used in the *actual* world is reliable. What counts as a "close possible world" (that is, what counts as part of the relevant reference class) determines the reliability of any particular method. Without an objective ordering of worlds, modal reliability is untenable.

Modal reliability, then, suffers from both the reference-class problem (from the missing closeness metric) and its special case, the generality problem of determining what exactly counts as a method of belief-formation. It looks as if we would be better off with a different understanding of reliability.

2. Probabilistic Reliability

An alternate way of understanding reliability is probabilistically: the reliability of a belief-formation method is the probability that a belief formed by that method is true. Unfortunately, probabilistic reliability *also* suffers from both the reference-class problem and the generality problem. Luckily, these concerns can be mitigated, at least in part.

First, let us consider the broader reference-class problem. Although this is traditionally only an issue in frequentist understandings of probability, Hájek (2007) has shown that all other informative theories of probability (classical, logical, propensity, and subjectivist) suffer from their own versions of the reference-class problem, and so a new (informative) interpretation of probability will be of no help. Intuitively, it seems like there should be some fact of the matter as to the probability of an event *A*. This first version of the reference-class problem is easily mitigated when we take seriously Hájek's insistence that it is *conditional* probability that must be primitive to a theory of probability. He has sufficiently shown that this must be the case (2003a, 2003b); the Four Horn Theorem shows unequivocally that the standard analysis of conditional probability as $P(A|B) = P(A \cap B)/P(B)$ for $P(B) > 0$ is simply untenable. Since we *do* have axiomatizations that take conditional probability as primitive (e.g. Popper, 2002; Roeper and Leblanc, 1999), it looks as if we can simply use one of *those* and we will be (partly) free of the reference-class problem. The simple answer to the reference-class problem is that there just *isn't* an unconditional probability "in the world" that is *the* probability of an event occurring. Instead, there are indefinitely many conditional probabilities.

That there are indefinitely many conditional probabilities "behind" any particular unconditional probability gives rise to the second form of the reference class problem: *which* conditional probabilities are the important ones? For example, if I want to know if my dog will

live to the age of 15, I inquire as to the probability of that event. I know that this is some complex of conditional probabilities: given some condition (or conjunction of conditions), I can learn the probability that she will live to 15. Do I look at all dogs? Do I look at all mutts? All black mutts? All black mutts with a particular pedigree? All black mutts of a particular pedigree named "Robot the Rocketdog"? Which conditional probabilities are relevant to the question? If I conjoin too many conditions, I end up with a sample size of one and inquire as to the probability, essentially, of my dog living to 15 given that she lives to 15 (or doesn't). To recall the question as to Smith's reliability in Barcelona Brown, what are the conditions for determining the probability of his belief-formation method getting things right? If "Brown is in Barcelona" or "Jones has a Ford" are part of the conditional probability, then he's quite reliable. Otherwise, Smith does not do nearly as well. Clearly, this will not do.

Goldman's distinction between global and local reliability will not help here. Local reliability is what we get when we are only concerned about the particular belief in question. The process that generated it is locally reliable just in case it produces *that belief* reliably. The process is globally reliable just in case it produces *beliefs in general* reliably. While this sounds a bit strange for, say, a frequentist probabilistic approach (the probability of getting a *particular* belief right just *is* the probability of getting other beliefs right), it makes more sense in the context of Goldman's modal understanding of reliability: local (modal) reliability is when we center the worlds in question on the conjunction of a belief-formation process and a particular belief that it generates, and modify the scenario in such a way as to potentially alter the veridicality of that process's result. Global (modal) reliability, however, centers worlds on the *process*, and looks for some sufficiently high ratio of true to false beliefs to determine reliability. Put another way, local reliability asks "How could the formation of that belief have gone

wrong?" while global reliability asks "How could that *process* go wrong in general?" (1986, p. 45-50).

It is obvious that, however we center our possible worlds or divide our reference classes, both global and local reliability are at the mercy of this division. Any statement of global or local reliability – even given a particular belief-formation process - can have its reference class or modal distance metric gerrymandered in such a way as to provide examples of its being both too strong and too weak for whatever application is at hand (see e.g. Pritchard, 2008; Comesaña, 2005; and Greco, 2007). The global/local distinction is unhelpful in this case, because the real, substantive problem of reliability ignores the distinction.

What, then, can we do? The first step is to not worry too much. The reference class problem is a problem in probability theory, but similar versions can be found everywhere. Hájek adroitly avoids the metaphysical reference class problem ("What's the right probability for this event?") by noting that *conditional* probability is the proper primitive concept for a theory of probability ("It just doesn't exist – all probabilities are probabilities given some condition or another"). This move will not, however, help with the epistemic reference-class problem ("Well, *which* conditional probabilities are relevant to my question?"). But this sort of question is everywhere. Many of our standard predicates suffer from sorites paradoxes, and these are something like reference-class problems for "heap," "bald," and so on. Arithmetic suffers from non-standard interpretations of the Peano axioms with no way to adjudicate between them – this is something like the "reference-class problem" of natural numbers. My account of belief from Chapter 2 makes it difficult to see, for a given set of actions, whether one belief or another is the best explanation. Even *truth* has reference-class problems, since we do not know to what degree our current best theories describe reality, and we have no idea what the full descriptive truth

contains (Barrett 2003, 2008). In all cases, we are simply doing the best we can with the tools we have. The reference-class problem (and its related difficulties in all fields of philosophy) is not cause to throw out a theory, but rather cause to *remain fallibilist*. It is simply *true* that our reference class substantively *matters* for determining probabilities. The best we can do is to describe a reference class for our particular application and hope that our interlocutors indeed agree that it is an appropriate reference class for the issue at hand. The reference class is stipulated as a working hypothesis, and it can be debated, altered, and indeed rejected. I propose that we go with frequentist probability because its reference class problem is the most obvious, and the familiarity of philosophers with its reference-class problem is widespread. There is simply no question that, when we understand probability in a frequentist way, the reference class determines our outcomes. It is *flatly obvious*. It is not that theory is substantively *better*, but that it shows its axiomatic assumptions vis-à-vis the reference class up front *as* assumptions to be discussed. It obviously and clearly embeds itself in ongoing diachronic inquiry, as opposed to appealing to innocuous-sounding but deeply problematic locutions like the notorious "close possible world."

If all these theories of reliability suffer from the reference class problem, though, why go for probability at all? Why not stick with modality? After all, philosophers have been comfortable with modal notions for decades (see, for example, Lewis 1973 and 1986; Stalnaker 1976, 1980, and 2003; Ellis 2005; Lange 2005; Handfield 2005; and Plantinga 1970 and 1974). My reasons for this are the following: First, as a matter of historical fact, discussion about close possible worlds has generally failed to make explicit what closeness metric or criteria of similarity among worlds is being used. The examples are myriad: Comesaña (2005), Nozick (1981), Sosa (2009 and 2011), Pritchard (2008), Greco (2007), and so on. Thus it seems better to

use a metric that makes its flaws more obvious, so that perhaps more philosophers will heed Hájek's plea to "join [him] in the search for a solution for the interpretations of probability that have a genuine claim to being guides to life" (2007, p. 84) instead of flatly ignoring the problem. After all, the more our assumptions regarding the reference class are made explicit, the better we – as a community of fallibilist philosophical inquirers – will be able to identify the salient aspects and begin to develop guidelines for the selection of appropriate reference classes and, one would hope, eventually an axiomatization of conditional probability that defines the appropriate reference class in an informative way.

My second reason for adopting a probabilist understanding of reliability is more pragmatic and aesthetic: modal logic is a good tool. We have a good understanding of the box and the diamond, and supervaluational semantics give us a fine way to talk about the box-arrow and diamond-arrow. Pragmatic and nomological possibility, though, require a dilution of this beautiful logic, and require introducing problems into it that greatly diminish its utility as a tool of analysis. We *already* have a theory that is fraught with assumption and difficulty, and that theory is probability. I see no compelling reason to introduce the problems of probability theory into modal logic. It is better to have a fine tool and a rusty (if serviceable) one than two rusty tools. When we consider that our everyday utterances of counterfactual conditionals (both in their necessitative "If *A* had happened, *C* would have happened" and possibilistic "If *A* had happened, *C* might have happened" forms) are equally well-expressed by probabilistic sentences ("If *A* had happened, *C*'s probability would have been high" and "If *A* had happened, *C*'s probability would have been non-zero" respectively), one wonders why modal logic was imported into the discussion at all.

3. Solving the Generality and Reference-Class Problems

Tu quoque responses do little to solve the issue at hand. Luckily, there is more that we can do with respect to the particular problems of generality and reference class with respect to belief-formation processes. When it comes to the reliability in this context, we have some fairly stable guidelines that help us determine both the precise method of belief-formation *and* the relevant reference class. I shall address the more specific generality problem first; I will then discuss the more general reference-class problem.

The generality problem is the problem of figuring out just what belief-formation method was used to generate a particular belief. There are indefinitely many types of belief-formation that could apply to any token belief equally well. Goldman suggests that the appropriate type is the "*narrowest* type that is *causally operative* in producing the belief in question" (1986, p. 50). But we have little reason to think that there would not be ties for "narrowest," and Baumann is right to point out that the requirement for a causal – as opposed to probabilistic – relation is unclear, and that there seems to be little reason as to why a narrow process is better than a more broad one (particularly in cases where the broader type is more causally efficacious than the narrower type) (2009, p. 84).

When it comes to belief-formation processes, humans believe according to what fits most coherently with their previous beliefs, including their beliefs about the reliability of any particular source of information. I suggest that *this* is the appropriate method of belief-formation to be referenced when assessing reliability. The process of coherence – generally described as a problem of constraint-satisfaction – is a development of the habit sketched in Alston (1995), and insofar as our best cognitive science understands how we form new beliefs, we know the general belief-formation process used. To the extent that a particular coherence network is likely to

produce true beliefs, it is reliable. Indeed, since our best science has a reasonably robust understanding of belief-formation in this capacity, we can sensibly define the relevant belief-formation method as *coherence with an agent's network of beliefs*. Thus the reliability of a belief's formation is identical to the reliability of the network of which it is a part. The reliability of a belief network is the probability that an arbitrary belief is true, given that it is coherent with the belief network in which it is situated. The generality problem is not a real problem, because the method by which we form beliefs need not be individuated beyond "*b* is a belief" and "Alejandra did something-or-other (it matters not what) and ended up believing *b* based upon its coherence with the rest of her beliefs." Alejandra will believe her belief with whatever credence she thinks is appropriate – we only need to ask how right she was in the end, *not* what her precise method was. Since we know that humans believe according to coherence considerations, coherence with other beliefs seems to be best description of belief-formation available.

This solution to the generality problem invites an easy solution to the reference-class problem for beliefs: the appropriate reference class is *the agent in question's beliefs*. To determine an agent's reliability (and therefore the reliability of the formation-processes of any of her beliefs), we simply inquire as to the mean degree of truth in their beliefs. If she believes many more false things than true, she is unreliable. If she believe such that most of her beliefs are easily recaptured in the language of the full descriptive truth, then she is reliable. We do not *need* to look outside her, because it doesn't *matter* what the external source of her beliefs was, nor does it matter if she used deduction, induction, abduction, or guessing. She forms beliefs according to *her estimate* of the reliability of her environment, herself, and her tools (including whatever she takes to be her current belief-formation method). It is irrelevant whether any particular external tool is reliable or unreliable, because her best understanding of its reliability is

reflected in her beliefs and their degrees, and if she estimates correctly then she will be more likely to form true than false beliefs (and will indeed have at least one true belief, about the reliability of her tool). We do not *need* an analysis of reliability external to an agent because all the data we could want to determine the reliability of her belief-formation process is right there, in her head.

The reference-class problem, then, isn't too much of a problem. If Alejandra has many more true beliefs than false beliefs, *she herself* is reliable. The local belief-formation method she uses is best-described as *coherence with Alejandra's beliefs*. The reliability of her beliefs is precisely determined by the veridicality of her belief-set, because she is constantly processing new data and doubting (and therefore inquiring about) beliefs that are problematic. Every belief in her head has her seal of approval, because otherwise she would *believe something else*, and so the reference-class problem (for the question of beliefs) is solved. The reliability of a local method is the reliability with which belief-coherence produces true beliefs in a particular agent. The reliability with which belief-coherence produces true beliefs in an agent is determined by nothing more than the truth of her beliefs. The various interpretation of the probability axioms are still problematic, but their problems can be held at bay when bringing the theories to bear on epistemology.

4. Epistemically Unfriendly Environments

There is one final worry with respect to reliability that I have left out: what happens in an epistemically unfriendly environment? There are two ways to solve this problem. The first is to simply state that the world *is* epistemically friendly, because "reality" is defined as those things

that are amenable to inquiry. If world *isn't* that way, then whatever facsimile, simulation, or what-have-you with which we *do* interact is "reality," and the man behind the curtain is simply something else, call it perhaps "the supernatural" (Peirce, 1878). But this first way is unsatisfying. We can imagine someone being raised on a television set *a la The Truman Show* and despite his eventual escape and discovery of the real world, it seems as if there is a sense in which the he is reliable, despite being massively deceived.

Enter the second answer. Recall Hájek's result: *conditional* probability is the proper primitive for a theory of probability (2003a, 2003b). Above, I was careful to say that the reliability of a belief-formation method (our use of coherence in particular) is the probability of producing a true belief *given* that we have the beliefs and commitments we do indeed have. There is a bit more to this, though, for the appropriate conditional probability should also include some degree of epistemic friendliness. Of course, we generally assume the epistemic friendliness of our environment in the first place, and so epistemic friendliness certainly counts as part of our beliefs and commitments. Nevertheless, it is useful to highlight that we are so committed.

That we are committed to a certain degree of epistemic access to the world is not so strange. Such access, after all, is a precondition for inquiry as such is necessary if we are to expect coherence to yield true beliefs at all. Shogenji's proof that repeated instances of coherence among independent sources can generate reliability from scratch (2005) fails if it turns out that what we thought were red balls pulled from the urn were really rabbits all along. For the process of diachronic inquiry to have *any* hope of being truth-seeking in the long term, and for coherence to *ever* reliably produce beliefs, we require an environment in which there is not radical epistemic opacity that we cannot mitigate. Thus we should understand the reliability of

Alejandra's belief network to be the probability that an arbitrary belief is true *given* that it is coherent with the rest of her beliefs. Among these beliefs, crucially, are (perhaps implicit) beliefs regarding the epistemic accessibility of her environment, the reliability of her tools (both psychological processes like deduction and external tools like computers), and so on. Our best way to calculate this probability is by looking at the truth of her current beliefs. In cases where this the precondition of epistemic friendliness is violated, then Alejandra is not reliable. We cannot fault her for this, of course: she is radically deceived by her environment. In such circumstances, it difficult to see how any of the epistemic virtues, the possession of reason, or anything else could be considered epistemically meritorious if there is no relation at all between possessing them and the ability to form true beliefs about the world.[2]

To sharpen the point, consider again the familiar case of Henry and his fake barns, modified from Goldman's (1976):

Barnsylvania: Henry is driving through the countryside. He sees a barn and judges that it is a barn. Unbeknownst to him, some high proportion of local barns are really barn-facsimiles. Although Henry has identified a real barn, if he had set his gaze upon a fake barn, he would have judged it to be a real barn. Was Henry's belief that he saw a barn reliably-formed?

It looks as if Henry is unreliable. In Barnsylvania, Henry's beliefs regarding barns, the reliability of his sensory faculties, that things are generally as they seem, and so on are false. Henry is particularly unreliable in Barnsylvania, and this unreliability is reflected in his belief-network: his indexical implicit beliefs (e.g. "Things are as they appear here, now") are simply *false*. Since the reliability of a belief network (or indeed of an agent) can be expressed as the proportion of true beliefs to total beliefs in that belief network and Henry has many false beliefs

[2] And indeed, Sosa tells us that "nothing is a cognitive virtue unless it is a truth-conducive disposition." (2011, p. 135).

about his environment, his reliability is quite low. However we understand the relation between justification and reliability, Henry is not reliable in Barnsylvania.

Of course, Henry is only temporarily unreliable. It is part and parcel of the example that Henry's situation is somewhat like the victim of a brain-in-a-vat scenario, and that he has no access to the truth about the barns (and that his belief network contains a large proportion of false beliefs about them). But of course this is silly. If one *really were* in a place like Barnsylvania, one would have all sorts of ways of finding out that it is full of fake barns. Moreover, it seems hasty to attribute such improbable beliefs as "The world is (perfectly) epistemically friendly" to Henry, when most reasonable people understand that they are occasionally wrong.

The most obvious reference class for determining Henry's reliability is the class of his beliefs. But presumably, very few of his beliefs are actually about barns. As such, insofar as his he has a large number of (perhaps implicit) true beliefs about the world, a relatively small number of false beliefs about barns need not doom Henry to unreliability. Without a fully-specified network of beliefs complete with truth-values, we cannot read off Henry's reliability from the example. In the most extreme case – where Henry's environment (in addition to having fake barns) is radically epistemically unfriendly (e.g. evil demon scenarios, brains in vats, and so on), then he is certainly unreliable. In cases where his access to barn-information alone is restricted but the world is generally epistemically friendly and Henry's belief network is generally veridical, then Henry is probably reliable, since coherence with his other beliefs is such a good indicator of truth most of the time. In the most realistic case where Henry is a normal person in a strange part of the world, he should be able to more or less quickly figure out that Barnsylvania is full of fake barns. From the point of view of the truth of the situation, Henry

should be able to point out the error of his previous beliefs. In the meantime, he is most likely generally reliable, though he has a few false beliefs about local barns. Insofar as justification depends upon reliability, Henry seems to be reliable precisely insofar as he is capable of revising his previous mistaken beliefs. But crucially, being in an environment where one can revise their mistaken beliefs just *is* being in an epistemically friendly environment. The inverse relationship between Henry's reliability and the epistemic unfriendliness of his environment is most obvious when we consider that if Henry is able to identify and correct his false beliefs about barn facades, then it becomes immediately plausible that perhaps he *was* reliable, after all. It is only on a presumption of complete epistemic unfriendliness – where Barnsylvania is something closer to a brain-in-a-vat scenario – that Henry begins to look like he is an unreliable sort of fellow.

In the case of the massively-deceived individual described at the beginning of this section, whether or not he is reliable is dependent entirely upon the proportion of true to false beliefs that he possesses. Insofar as he generally has true beliefs about the world – insofar as coherence has guided him towards the formation of true beliefs – he is reliable. If his deception is particularly deep, then it might be the case that his ratio of true to false beliefs is not high enough to count as "reliable." In either case, his reliability is inversely proportional to the depth of his deceit, that is, to the epistemic unfriendliness of his environment.

5. Conclusion

I have argued that epistemologists need not worry about the generality problem of reliabilism. While the best understanding of reliability we have – probabilistic frequency – is not perfect, it is serviceable for the case at hand. In the case of individuating a belief-formation

process, the appropriate process is coherence with an agent's beliefs. To the extent that an arbitrary belief is likely to be true if it is coherent with an agent's beliefs, that belief is reliable. In determining the probability of an arbitrary coherent belief's being true, I suggest that the relevant reference class is the set of an agent's beliefs. To the extent that an agent has true beliefs, she is reliable. Insofar as belief itself is well-understood, epistemologists can feel free to make use of the concept of a reliably-formed belief in explicating knowledge, justification, and epistemic virtue.

CHAPTER 6: KNOWLEDGE

In Chapter 2, we saw that the sense of truth that is relevant in the locution "true belief" is not a matter of full, descriptive truth. If we are after that, then none of our beliefs are true. Instead, the truth in "true belief" is a matter of being able to recapture the "true" belief from the point of view of the full descriptive truth. To the extent that our "true" belief is maintained in and recognizable from the point of view of the full descriptive truth, it is true. But it is enormously likely that *none* of our beliefs now match perfectly with the full descriptive truth. Even simple sentences like "Dogs are mammals" are liable to change somewhat in their meanings. Dogs (and mammals) are living, which means that they are organic, which means that they contain carbon, which is an atom, which have electrons. And we do not know the full story about electrons. We are quite happy to allow vagueness about electrons into "Dogs are mammals," but allowing the vagueness is just admitting that it is there. "Dogs are mammals" is true *enough*, and say what you want about electrons, we are committed to being able to understand "Dogs are mammals" on contemporary lips from the point of view of the full descriptive truth. But saying *that* is not the same as saying that "Dogs are mammals" is a true to the maximal degree. It is in fact *denying* that "Dogs are mammals" is true in the sense of the full, descriptive truth. It just so happens that the full, descriptive truth is not so interesting – it is more important that the sentence is true *enough*, or approximately true. The degree to which the sentence is true – the degree to which it is a primitive articulation of the full, descriptive truth – is its T-measure.

Just as the truth we care about (though not the full descriptive truth) comes in degree, belief itself comes in degrees and in a number of senses. The first sense is just that we can

believe things more or less strongly. In an ideal agent the distribution of degrees-of-belief would match with some calculus or another (Kolmogorov's probability calculus if one is a Bayesian, a possibility calculus that is somewhat similar to fuzzy logic if one considers the ability to accommodate vagueness important to a calculus of belief). This is easy to understand and uncontroversial. The *other* way that belief comes in degrees, though, is in how *belief-like* a system's beliefs are. Taking something to have beliefs in the first place is taking its outputs to be best explained by an ABI model. Using the ABI model to understand something implies that there is an agent that performs *actions* to carry out *intentions* guided by *beliefs*. When we interpret a bonobo or a group or a dog's actions *as* actions, or interpret an orangutan's behavior as guided by *belief*, we apply the ABI model to them. But of course a dog-belief is different from a bonobo-belief, and both are different from my beliefs. Some things, like chairs, aren't belief-like systems at all. Many other things, though – most obviously groups of humans, higher mammals, and certain computer programs – seem belief-like, if not perfectly like humans. Such systems are belief-systems to some extent or another. The second sense of degree of belief is the degree to which the rules a system follows *count as* beliefs, the degree to which that system really is an ABI-system.

Truth comes in degree, belief comes in degree, and the applicability of belief interpretations to a system comes in degree. Justification, too, comes in degree. Synchronic justification – that is, coherence – comes in degree as described in Chapter 3, and it is simply a fact that humans use coherence in thought. Methodological justification – the degree to which a proposition is appropriately embedded in the context of ongoing fallibilist inquiry – comes in degree as described in Chapter 4. If knowledge turned out to be a binary predicate given that all its traditional components are matters of degree, it would be quite extraordinary.

1. Knowledge as Scalar

Knowledge is a complex of justification, truth, and belief. Human beings are such that we just *do* attempt to maximize coherence (that is, synchronic justification). So much the better for us. It is a further part of our nature that we seek to replace doubts with beliefs. Some of these beliefs are true, and some are false. While it is part of human nature that we really do inquire when occasioned by doubt, we can give a measure (as in Chapter 4) of how appropriately fallibilist an inquirer is. Of course, the "appropriately" in "appropriately fallibilist" is by our own current lights, and what counts as the "right amount" of fallibilism in inquiry is open to revision.

I take it that knowledge is indeed a matter of degree. It is something like the (perhaps weighted) product of the measures of truth, methodological justification,[1] degree of belief, and appropriateness of using the ABI-model. Each of these measures is revisable, and our current best bet for measures like "truth" and "applicability of the ABI-model" are fairly unrefined. Computer scientists have been making some progress in defining measures of conceptual similarity for machine learning, but such efforts are not robust enough to be usable for comparing entire conceptual apparatuses. And, of course, even if we *could* generate a measure of e.g. truth, it would be useless. We all believe what we take to be true. If we knew how our beliefs would be revised in the future, we would go ahead and revise those beliefs right away. Even if we had a measure of conceptual similarity, we would not have access to the full

[1] Synchronic justification can drop out here, as we get it for free as out methods become more and more reliable. As degree-of-belief and the J-measure trend up, coherence does as well.

descriptive truth to test a concept's similarity to it. This is rather distressing, because we can't know for sure whether we know a proposition. Instead, we commit ourselves to that proposition's preservation in future theories. When we say that a belief is *true*, we say that future revisions of our beliefs will recognize that belief as an approximation of the way that the world really is.

Taking knowledge as a scalar measure has an obvious benefit: strange cases like those discussed in the previous chapters simply evaporate. All believed propositions are candidates for knowledge in some sense. Those propositions that are completely wrong-headed (and so are not approximately true to any appreciable degree) or categorically denied (and so have an extremely low degree of belief) or which are believed for silly reasons without sensitivity to the unreliability of those reasons (and so have extremely low measures of methodological justification) have extremely low degrees of knowledge. Of course, just what counts as a high enough degree of knowledge for knowledge-attribution in a particular context is imprecise, and so my account of knowledge lacks the finality of an analysis along the lines of necessary and sufficient conditions. But the world is a complicated place, and it does not seem unduly surprising that necessary and sufficient conditions – perhaps the simplest of a philosopher's analytic tools – are not nuanced or robust enough to adequately describe the particular complication that is knowledge.

Of course, the knowledge-as-scalar view might be too extreme for some. Because I simply lack the intuition that knowledge must not admit of degree, I am comfortable accepting that knowledge-that-*p* is a matter of degree, full stop. But if this is too much, it is easy to think of ideal knowledge as a case similar to full descriptive truth: just as *nothing* we say now is part of the full descriptive truth, nothing that we believe now is an item of ideal knowledge. On this

model, ideal knowledge is the mental state at the end of inquiry, when one has perfectly mapped the world through scientific inquiry. Because all the data is in and there are no doubts, we are certain in all our beliefs, and of course we could not be wrong. Because we have access to the full descriptive truth, the relevant propositions are part of the full descriptive truth. In short, every relevant component of knowledge is maximized at the end of inquiry, and so knowledge is as well. *That* is ideal knowledge, and the scalar I describe is a degree of its approximation. Just as the degree of truth in an uttered sentence is the degree to which it approximates (that is, can be understood as articulating the same thing as) the full descriptive truth, degree of knowledge in an uttered knowledge-attribution can be understand as the degree to which an ideal knower would be able to recognize the mental state in question as true, methodologically justified, and so on.

While such a view is unorthodox, it is quite advantageous in how it deals with skepticism and scientific pessimism. What, exactly, could the skeptic say that would bother us? It is no surprise that we lack perfect knowledge. He might tell us that our own estimations of how closely our beliefs approximate ideal knowledge are too high, but this, too, is not surprising in the absence of access to the full, descriptive truth. When belief is understood as a rule for action, it becomes apparent that insofar as the skeptic's arguments do not occasion genuine doubt and an observable change in action (of which private judgment is a species), we do not *need* a particular response to him. Indeed, because our actions bear out our beliefs, his insistence that we stop believing that which we believe can only be seen as a demand that we stop acting. And, of course, if we are to take the skeptic seriously, he must take himself seriously and stop acting as well. His own lack of action, though, defeats his position. Here too I agree with Barrett: "one cannot help but wonder at the sincerity of [the skeptic's] own claims to doubt" (2001, p. 13).

With a Peircian account of truth and belief at our backs, it begins to look as if the best a skeptic could do would be to give a genealogy of our beliefs in which the contingent factors of them – our being born in a specific time or a specific place, to borrow Barrett's example (2001, p. 8) – are made obvious. But the contingency of my believing *p* does not mean that *p* is false. Indeed, if I thought that there were good reason to believe ~*p*, I would do that – I believe *p* because *p* best explains the data that is my conscious experience of the world and conforms to the methodological commitments that I *really do* have and *really do* take to be my best bet for forming beliefs. I may not be able to articulate these commitments, but the skeptic's genetic story doesn't tell me *that* I am wrong. Instead, it tells me that if some things had been different, if I had been born in a different time and in a different place, or with different neurological structures, or with a different attitude towards investigation, I might have gone wrong. Insofar as I am confident that I *really have* made the best decision regarding *p*, skeptical arguments can do nothing to convince me that ~*p*.

A Peircian treatment of scientific pessimism runs along largely the same lines. A pessimist tells us that induction on the local success but ultimate falsity of past scientific theories gives us good reason to suppose that our current theories are false (Laudan, 1981). The Peircian fallibilist, of course, fully *expects* that our current theories aren't part of the full descriptive truth without retrospective translation. But what of this? We know that the sentences of our current scientific vocabulary probably do not mean what we think they do, but we also expect that certain existence claims, observational data, methodological commitments and so on will be preserved in the next generation of scientific theories. We expect that the next generation of theories will be able to explain not only how our current theories have gone wrong, but also the sense in which our current theories are *right*. Because scientific progress – much like human

interaction with everyday medium-sized objects – proceeds according to fallibilist commitments, the Peircian neither expects nor attempts to believe a proposition with absolute certainty. All sentences are hypotheses. We hold on to some of these sentences and sentence schemata very strongly. Apparent violations of noncontradiction, for example, are more likely to occasion us to doubt our data than our logic. But *certainty*, identification of *any* current sentence with the full descriptive truth, is precisely an abandoning of fallibilist inquiry as such, and so the Peircian can only question why anyone would want to embark on such a project in the first place.

2. Epistemic Luck

According to traditional analyses of knowledge, epistemic luck is a problem: luck defeats knowledge. Hopefully, the previous chapters have made it clear, though, that this interpretation of luck is simply untenable. *All* our beliefs are lucky to some extent. We can be quite confident that none of our beliefs are part of the full, descriptive truth. *All* of our beliefs would require some charitable reinterpretation to account for the way that the world really is. The same can be said for our belief-forming methods. None of them are perfectly reliable, since they cannot even produce a single completely true belief. Insofar as we have *any* true beliefs, we are *lucky* that they are reinterpretable from the point of view of the full descriptive truth. We are *lucky* that we went with Copernican astronomy instead of further revisions of the Ptolemaic model. We are *lucky* that no one came up with a model of phlogiston or caloric that fit the data. If *any* of our scientific beliefs are true (or true enough), then we are lucky that they are true (or true enough) because we are *currently* working with an inconsistent system.

The same can be said of everyday beliefs. When I ask a stranger for the time, I'm lucky that she doesn't play a joke on me. I'm lucky that she doesn't have a broken watch that always reads 4:00pm. I'm lucky that she doesn't have such a watch even when it actually *is* 4:00pm. Even when my belief isn't due to an inordinate amount of luck, I'm lucky that it isn't. An infinite number of things could change that would lead to me having different beliefs than I have. Whenever I believe something, I am lucky that I'm not wrong. I'm lucky that the general categories into which I place everyday objects are actual features of the world (or approximate the actual features of the world to a reasonably high degree), and not silly fictions.

By my lights, luck is inescapable and is suitably accounted for by understanding knowledge as scalar. While I therefore take it that traditional responses to luck in general and to Gettier cases in specific have therefore missed the mark, the problems that traditional models of knowledge seek to answer and eliminate are indeed genuine problems, and the intuitions they encode offer us deep insight into knowledge. To use the vocabulary of diachronic inquiry, past theories of knowledge have been tremendously successful, and any subsequent theory of knowledge must make sense of their success. I begin this account with Engel's identification of two types of epistemic luck: evidential and veritic (1992). We are evidentially lucky if we're just lucky to have the evidence that we do have. We're veritically lucky if, given our evidence, it's lucky that we're right. Engel takes evidential luck to be compatible with knowledge, but takes veritic luck to be incompatible with it. Various writers have, since then, understood veritic luck in various ways: popularly, veritic luck might violate some modal condition on knowledge such that our belief in *p* must covary with *p*'s truth across a range of possible worlds. Call such positions *covariantist*.

But covariantism suffers from the flaw that it makes use of the inscrutable notion of the closeness of possible worlds. Matters are made worse when the method of cases is used to argue in favor of the various modal conditions – what is really being argued is, of course, the proper outputs of the closeness function that orders the possible worlds. Two worlds can be made to be arbitrarily close or far from one another by varying the closeness metric; since the closeness metric is rarely if ever even *mentioned* in modal epistemology, we would do well to take the cottage industry not as a discovering facts about *knowledge*, but as offering various understandings of *modal distance*.

There is one notable exception to the above worry: Sherrilyn Roush argues that knowledge is belief that tracks the truth. She uses the sensitivity and safety conditions. Rather than wandering down the garden path of close possible worlds, she defines > probabilistically: $A > C$ iff. C obtains in at least 95% of the cases in which A obtains. If A is such a strange and remote condition that we have *no idea* what to say about C, such that we cannot say whether or not it happens in 95% of the cases, then the scenario was malformed in the first place (thus ending the debates about whether I would know "The party is at such-and-such location" if I were going to Andy's party, and Andy hired Judy to give directions and also to call him to move the party if she saw Michael, and I almost disguised myself as Michael but decided not to at the last moment, and so eventually received correct directions from Judy) (Comesaña, 2005).

What Roush clearly articulates – and what other philosophers who make use of modality hint at – is the intuition that knowledge tracks truth. We believe p in all and only those situations where p is actually true. Thus if our method is off, or we believe for the wrong reasons, or one of any number of wrinkles appears such that we would end up believing $\sim p$ when p or *vice versa*, it begins to look like we don't know p after all.

But this is too strong. Surely a single logically possible but utterly ridiculous scenario (if I had been looking at a fake zebra I have been fooled into thinking it was real; I'm looking at a real zebra, but I still don't know it) does not remove knowledge in the average case. This is what I take to be essential to Roush's program: knowledge allows for the occasional mistake. Modal understandings usually assume (tacitly or otherwise) a Lewisean semantics of the counterfactual conditional, but Lewis's counterfactual conditional does not allow for exceptions. But knowing *p* doesn't imply that we are infallible at determining *p*. If we try to limit things to "close" possible worlds, this just means that we are infallible in "close" *p*-worlds, and so any argument as to whether I made a mistake can instead be shifted to an argument about modal distance.

The fundamental intuition behind covariantism is the notion that the knower is appropriately responsive to the truth and falsity of the known proposition. Some philosophers abandon the understanding of this intuition that involves possible worlds in favor of a new primitive term: knowledge is *reliable*. This position is, quite reasonably, called *reliabilism.*

Peter Baumann (2009) rightly rejects modal notions of reliability due to the missing closeness metric. Probabilistic notions of reliability suffer from the difficulty of determining the relevant reference class. For example, in the case of Barnsylvania, do we include the probability of making an incorrect barn-attribution for any barn in the town of Barnsylvania, for any barn in the state, or any barn in the world? If we restrict the reference class too far (perhaps to a single case), then we lose relevant statistical information (which is particularly important if we are frequentists about probability). Where modal accounts of reliability lack the ability to determine modal distance, probabilistic accounts lack the ability to determine the size of the class from

which probabilities are determined.[2] To solve this, Baumann proposes that the reference class is determined by context, intent of the speaker, and so on.

I think that Baumann is right (and insofar as reliabilism is right, the underlying implied covariantism is right, as well). Knowledge should be reliably-formed. The bar with respect to *how* reliable it must be varies across contexts. And indeed one could characterize the understanding of methodological justification described in Chapter 4 as a measure of a method's reliability. More precisely, the measure from Chapter 4 is a measure of the error with respect to how an agent develops his or her beliefs according to the reliability of the particular (local) method being used. A high measure of methodological justification means that reliable belief-forming processes produce beliefs that are more tenaciously-held than those produced by unreliable processes. The net result of this is that a belief-system with a high degree of methodological justification *itself* becomes a reliable system: it only accepts beliefs that increase the overall coherence of the belief-network, and is *ex hypothesi* more likely to reject false beliefs than it is true beliefs. Thus coherence – the metric which underlies human reasoning according to our best contemporary science – is implemented by humans in such a way that the underlying belief-systems that instantiates coherence is reliable in proportion to the extent to which it is methodologically justified. Over repeated application, coherence yields reliability.

Given this result, it would be a mistake to declare the reliabilists wrong. They are, after all, correct in a way: reliability is an important component of knowledge.[3] Virtue epistemologists, too, are right in their own way. Wayne Riggs characterizes willingness to seriously entertain new beliefs – that is, taking oneself to be fallible – as *open-mindedness*

---

[2] This, of course, is a problem for Roush as well.
[3] As it turns out, reliability is an outcome of a belief's having both a high degree-of-belief and a high J-measure.

(2010). This sort of fallibilism is, of course, a precondition for methodological justification. If we understand open-mindedness as the ability to accept new data in the right circumstances (that is, to think of oneself as fallible when one really *is* wrong or likely to be wrong), it is natural to understand *intellectual courage* as the ability to reject new data under the right circumstances (that is, to think oneself correct when one really *is* correct or likely to so be). The combination of the two is having the ability to reject beliefs when appropriate (i.e. when one is wrong or the belief was formed unreliably) and to cling to beliefs when appropriate (i.e. when one is correct or the belief was formed reliably). This just *is*, by definition, that which Chapter 4's J-measure quantifies.

To this end, it would be incorrect to broadly declare virtue epistemologists wrong. After all, I agree with them that an item of knowledge is produced by belief systems that have particular properties. When these properties are reliable for getting us true beliefs, we tend to call them *virtues* and when they are unreliable, we tend to call them *vices*. Greco's thesis that knowing *p* is succeeding in believing the truth about *p* because of intellectual ability is, of course, true (Greco 2010, p. 71). But it is not particularly illuminating: a belief system's having an "intellectual ability" in the relevant sense is having a stable property such that it reliably finds the truth with respect to some domain. While I have no desire to contradict the virtue epistemologist's claim that virtue is a necessary component of knowledge, this is because "virtue" is the term we use for a stable property of a belief system that causes it to be reliable.

Bernecker (2011) champions a position he calls "identificationism," wherein an agent knows that *p* if she truly believes *p* and her reasons for believing *p* either identify or are causally connected with *p*'s truthmaker. And indeed, we should expect diachronic inquiry to produce beliefs that correctly identify a proposition's truthmaker. Bernecker's position is better than

most, because it avoids the problems inherent with modal epistemology and the problem of the relevant reference class, but allows enough vagueness for us to make sense of knowledge in the absence of access to the full descriptive truth. What one takes to be the truthmaker of a proposition merely has to "have a lot to do" (p. 142) with that which actually makes the proposition true. For example, when we misidentify a wolf as a dog and then infer "That is an animal" from the false premise "That is a dog," we still know "That is an animal." After all, being a wolf and being a dog are pretty close to the same thing. From the point of view of the truth of the matter, we can see that thinking "That is a dog," while not *true,* gets enough of the relevant features of the situation right that it can grant a relatively high degree of knowledge. We are comfortable saying that dogs and wolves are similar enough that inferring from *either* of them to propositions about zoological classification is reliable. And moreover, the *actual* reliability of this process is high enough that beliefs formed by methods with a high J-measure are strongly-held. Thus we think that dog-based beliefs basically identify the same features of the world as wolf-based beliefs when it comes to attributions of the property "is an animal."

While the identification in "identificationism" is impossible without access to the full descriptive truth, Bernecker's allowance that there need only be some kind of causal connection between my reasons for believing *p* and *p*'s truthmaker is apt. Diachronic inquiry, after all, is a series of interactions with the real world wherein failed action generates doubt. Indeed the core of diachronic inquiry is that an inquirer can tell a story – a causal story – about why she believes what she does, about how she got to her current beliefs from her old beliefs.[4] This story is precisely the story of how her old beliefs failed in action and how her new beliefs account for

[4] This is not to say that we remember the full genealogy of all of our beliefs. Humans do forget things, after all.

those failures (that is, how they succeed where old beliefs failed) and how they subsume the successes of her previous beliefs. This story is a history of causal interactions between an inquirer and the external environment around her (which, of course, includes *other* inquirers). What's more, this is *precisely* the kind of genealogy that we expect that we should be able to tell from the point of view of the full descriptive truth about our current beliefs. To the extent that our current beliefs are maintained in that story told from the point of view of the truth – to the extent that our current beliefs are not dead-ends of inquiry and serve as rough but recognizable forms of the *real truth* in the genealogy of the full descriptive truth – we can say that Bernecker's desired causal relation is present. Our beliefs are causally related to the truth insofar as they are part of a genealogy of ideas that eventually will hit upon the truth. Thus Bernecker's relation is causal, because inquiry is a sequence of causal interactions between us and reality. But mere causality is not enough – the relation between what we *take* to be a proposition's truthmaker and that proposition's truthmaker must be one of approximate truth to whole truth (there are, after all, all sorts of silly stories where there is *some* causal connection between our reason for believing *p* and *p*'s truthmaker). The former must be recognizable as a primitive articulation of the latter. By understanding our beliefs as embedded in truth-seeking diachronic inquiry, we see precisely what sort of causal connection is required between reason and truthmaker: our reason for believing *p* must be recognizable from the point of view of the full descriptive truth as suitably relevant to *p*'s truth.

The "suitable relevance" of the preceding paragraph is extremely important, and in fact does the work of Bernecker's original use of the causal relation (as opposed to wholesale identification). Imagine someone who believed "There will be a full moon tonight" on the basis of "It is low tide." There is *some* causal connection between the two, since both are determined

by the moon, but our silly individual probably does not know that there will be a full moon tonight, even if it is true. While there is *some* causal connection between the two things, whether or not it is low tide right now does not make a *difference* as to whether or not there will be a full moon tonight. There is a causal connection between the two, but the former is not a causal difference-maker to the latter. Michael Strevens (2006) would say that the former does not *explain* the latter. Once we begin using the language of explanation, it becomes quite natural to embed the entire practice into the process of diachronic inquiry.

Put another way, at the end of inquiry we have the full, descriptive truth about reality. The full descriptive truth is in a (causal) genealogical relation with previous theories. Insofar as we can recognize an explanatory relation – that is, a causal difference-making relation – between S's reason for believing *p* and *p*'s truthmaker, there is a second causal relation partially identified by Bernecker. In the case of those entities whose existence is not directly verifiable like numbers, we believe in them because we need to believe in them to make sense of reality. The process of diachronic inquiry gives locally acausal entities like numbers a type of causality: genealogical causality, such that the cause is the positing in a theory and the effect is changes in future theories (most specifically, the theory that accounts for reality, i.e. the full descriptive truth). This allows my understanding of methodological justification to make sense of the clear successes of Bernecker's identificationism while showing its limitations, particularly its inability to account for necessary truths and falsehoods.[5]

I understand covariantism and reliabilism to be articulations of the intuition that knowledge could not easily be wrong. I take identificationism to be an improvement upon these

---

[5] Bernecker's account is, of course, not unique in having difficulty coping with tautologies and contradictions.

theories, as it notices that the best way to be reliable (and thus satisfy covariantist requirements while avoiding the problem of the reference class and the generality problem) is to have your *reasons* for believing a proposition suitably involve what makes that proposition true. If that is the full story, though, we know nothing because we do not have access to the full descriptive truth.

3. Measuring Knowledge

Throughout the past five chapters, I have described understandings of the main components of knowledge: truth, belief, and two senses of justification. It is perhaps unfortunate that none of these are binary matters. Knowledge is a complex concept and the knowledge that we care about as part of local attribution, as relevant to local inquiry, and so on comes in degree, even if full-blown or ideal knowledge does not. In local cases, we know a proposition better when that proposition is legitimately a belief (and this the ABI model is maximally applicable), and when we strongly believe that proposition, as described in Chapter 2. To strongly believe a proposition in a (methodologically) justified way is to have formed that belief by a reliable method. One component of that method is our propensity to believe that which coheres with our current beliefs, as described in Chapter 3. The coherence component of human belief-formation and the human tendency to doubt in the face of failed action together ensure that inquiry, in the long term, is a truth-seeking practice. A naïve formalizations of degree-of-knowledge might be something like the following:

$$K = M \times T \times B \times J$$

Degree of knowledge (or, closeness to ideal knowledge) is equal to the product of the applicability of the ABI-model to the system in question (M), the degree to which the believed proposition is recognizable as part of the full descriptive truth from the point of view of the full truth (T),[6] the degree of belief in the proposition, and the degree of methodological justification of that proposition. In the most usual case, where we are talking about human beings that are obviously belief-systems, we can drop the first of the terms, M. Thus knowledge is degree of truth times degree of justification times degree of belief.

There is good reason, though, to think that B, degree of belief, should not be part of the calculation. It seems obvious that insofar as a high K-measure reflects epistemic merit, T and J should be high. In the event that we are talking about belief-like systems rather than assuming a system in question is amenable to interpretation as an ABI-model, it could be useful to include M as well. B, though, encodes rather precarious assumptions about knowledge.

The most obvious objection to including B as a component of K is that B is already considered in calculating J, and so it should not be double-counted. But, of course, J is a measure of B's optimality, not of its magnitude. Including B in K is a matter of whether one considers a more strongly-held belief epistemically better – more knowledge-like – than a less strongly-held belief, not of double-counting.

It simply does not seem to me that a more strongly-held belief is *ceteris paribus* better than a weakly-held belief. What is meritorious about a belief is its being *true* in the first place, and its being embeddable in diachronic inquiry in the second. To motivate this, consider that I believe that there is a dog under my porch with credence .80 (let us ignore T and M for the

---

[6] I have characterized this previously as "how true" the proposition is.

moment). The reliability of this belief's formation was .90. At present, the J-measure of my belief is .90, and it would also be .90 if I believed with credence of 1.0. We can discover the epistemically optimal thing to do by finding the maximum of $(1 - |.9 - x|) \times x$ (that is, $R \times B$). It turns out, then, that the best thing to do in this situation is to believe that there is a dog under my porch with credence of .95.

This result, though, is absurd! Surely I should not be epistemically "rewarded" for increasing a belief's credence in a way that decreases that belief's justification. A sensible measure of knowledge – whether we interpret it as *knowledge* or approximation of an ideal – should not privilege hard-headed tenacity over justified believing. Insofar as the knowledge-measure involves degree-of-belief, it is surely not a simple matter of multiplication.

The initial intuitive appeal of including B in K is explained away when we note that a maximal K-value – ideal knowledge – must assign B a value of 1.0 even when B is not part of K. If we assume that K includes J, and that J is 1.0 (this is ideal or maximal knowledge, after all) then $R_{Mb} - B = 0$. Crucially, reliability increases in accordance with approximation of knowledge. As discussed in Chapter 5, "reliability" is best understood as mean degree of truth of an agent's beliefs, since we all use the broad method of "believe what best coheres with my beliefs," and this method applies to all our beliefs. Thus in cases where we assume that one believes only true things – such as the ideal case of the end of inquiry – we can also assume that one is perfectly reliable. In the ideal case of knowledge – maximal knowledge, R_{Mb} is 1.0.

Accordingly, B must be 1.0 as well.[7] Ideal knowledge, then, implies certain belief and perfect reliability.

High credence, then, is indeed a mark of ideal knowledge. The problem is simply that it is *also* a mark of a low degree of justification. We can make room for the intuition by simply requiring that K is only defined when B is at or above a certain degree. Requiring that B > .50 is reasonable, as a credence of greater than .50 seems to be the minimum for the sense in which one can be said to act according to a *rule* that *p* as opposed to randomly (if B = .50) or according to a rule that ~*p* (if B < .50). This, though, is a thesis on belief-attribution more than knowledge – the credence at which it is appropriate to say that a person really *believes* a proposition is also the credence at which it is possible to say that she *knows* it, assuming of course that it is true and justified to an appropriate degree.

Accordingly, let us modify the K-measure to exclude B:

$$K = M \times T \times J$$

As degree of belief tends to one and the degree of justification tends to one, the reliability of the belief-forming method also tends to one. And of course as reliability goes to one, degree of truth goes to one. As Gettier-style veritic luck divorces our reasons for believing a proposition from that which grounds its truth, false beliefs about our surroundings, the world, and so on cause us to require more charity to reinterpret our current position from the point of view of the full

[7] Here one might entertain a Gettier-type case: would the K-measure not give an unsatisfactory result if we get lucky in our sentences? That is, even though we do not have access to the conceptual schema capable of articulating the full descriptive truth, is it not possible that due to some extremely odd circumstance, I occasionally say something that is part of the full descriptive truth? Even if I am incredibly wrong about what I *think* I mean by the sentence, it isn't beyond thought that I could get extremely lucky and produce a sentence with a truth-value of 1.0. One might thus take my argument above to imply that my luckily-true sentence was also produced with a reliability of 1.0 as well. This argument, though, only goes through when we can assume the truth of not just a single proposition, but of *all* propositions in a belief network. Similarly, because evaluating reliability involves evaluating the truth in the totality of our beliefs including beliefs about the truth-conducivity of our sensory apparatuses, tools of investigation, and so on, there is no way to luckily or accidentally use a reliable belief-formation method.

descriptive truth. That is, T drops as luck enters the equation, even if we (luckily or knowingly) assign credences appropriately, and so K trends down as well.

As K trends up T and the presence of luck decrease. Because the T-measure of a sentence involves all the concepts employed in articulating that sentence, and these concepts themselves are evaluated in the reliability measure, we can expect to see overall reliability – and thus B – increase as T increases.

While I take it that the properties of the K-measure I describe are good properties for it to have because they accord with our intuitions regarding knowledge-attribution, other formulae instantiate the very same properties and relationships. Understanding K as the mean of ABI-applicability, truth, and justification

$$K = \overline{M + T + J}$$

for example, or understanding M not as a component of knowledge but having a high (enough) M-value as a precondition *on* knowledge is perfectly acceptable. I take it that this work's contribution to our understanding of knowledge comes in its analysis of its components and in articulating the structure of knowledge (or, at least, empirical instances of it) as a scalar measure, not in describing the exact function that produces the K-measure.

An inquirer is free to choose between knowledge-as-mean and knowledge-as-product according to whether one desires its curve to be linear or quadratic respectively. Insofar as any particular application of a knowledge-measure is embedded in some context or other, it is not unreasonable that we might expect that context to imply specific coefficients or formulations of the measure, such that it is enormously context-sensitive. Thus if we are interested in, for example, creating a belief-system that is a good predictor of medical diagnoses given some set of

symptoms, we could determine the appropriate K-measure for the learning context (i.e. medical data) by allowing an evolutionary algorithm to select coefficients (presumably with at least two "species:" linear and quadratic K-measures) where fitness is considered correct prediction. Once an appropriate K-measure is determined, we would then use that K-measure as the success condition for a supervised learning algorithm, thereby creating an expert system with the ability to make diagnoses and assign that diagnosis an appropriate level of confidence. While I take it that more detail in this avenue would be too much for my present conceptual aspirations, the value of producing a measure of knowledge that can actually be measured, quantified, and reported and iterated upon will be made particularly obvious by its practical application.

4. Conclusion

A proper understanding of justification, belief, and truth gives us a clear-eyed way to account for the various intuitions that have guided epistemological inquiry in recent years. Covariantists get what they want: as K trends up, error in evaluating *p* is eliminated. Reliabilists are similarly satisfied, since the higher our degree of knowledge, the more reliable the method that produced our belief. Virtue epistemologists get what they want, because those properties of a belief-system that allow it to produce knowledge can be rightly called "virtues," enumerated, and even taken as the primary object of interest in a theory of knowledge. The cutting edge – identificationism – gets what it wants. Coherentists are vindicated, because our best science tells us that coherence is fundamental to human cognition and, as mentioned, as justification, truth, and degree-of-belief trend up, coherence comes along for free as a result of human biology. Even the contextualists are shown to be correct, since what K-measure counts as "good enough"

for any particular attribution of the common binary predicate is surely (as with most words in natural languages) determined by the context of both the speaker and the subject (though I will leave complete analysis of *that* to other philosophers). The only epistemological theses that *aren't* vindicated in some way are those that require justification to be a wholly linear matter, and that is not the fault of my theory of knowledge but of the facts of the matter regarding how humans reason. Understanding each of knowledge's components not as binary predicates, but as measures that admit of degree is vital to understanding knowledge itself. Once we have done so, it turns out that Gettier's famous question can be answered in the affirmative: knowledge is justified true belief after all.

REFERENCES

Aaronovitch D. (2009). *Voodoo Histories: The Role of the Conspiracy Theory in Shaping Modern History*. London: Jonathan Cape.

Alston, William P. (1995). How to think about reliability. *Philosophical Topics* 23 (1):1-29.

Barrett, J. A. (2001). Towards a pragmatic account of scientific knowledge. Retrieved from http://philsci-archive.pitt.edu/id/eprint/472

Barrett, J. A. (2003). Are our best physical theories (probably and/or approximately) true? *Philosophy of Science* 70 (5):1206-1218.

Barrett, J. A. (2008). Approximate truth and descriptive nesting. *Erkenntnis*,*68*(2), 213-224.

Baumann, P. (2008). Is knowledge safe?. *American Philosophical Quarterly*,*45*(1), 19-30.

Baumann, P. (2009). Reliabilism—modal, probabilistic or contextualist?. *Grazer Philosophische Studien* 79 (1):77-89.

Bencivenga, E. (1994). *Philosophy in play: Three dialogues*. Hackett Publishing.

Bernecker, Sven (2011). Keeping track of the Gettier problem. *Pacific Philosophical Quarterly* 92 (2):127-152.

Bishop, Michael A. (2002). The theory theory thrice over: The child as scientist, superscientist, or social institution? *Studies in History and Philosophy of Science* 33 (1):121-36.

Bouvier, Alban (2007). Collective Belief, Acceptance, and Commitment in Science. *Iyyun* 56:91

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: the generic conspiracist beliefs scale. *Frontiers in psychology*, *4*.

Cao, Tian Yu & Schweber, Silvan S. (1993). The conceptual foundations and the philosophical aspects of renormalization theory. *Synthese* 97 (1):33 - 108.

Chisholm, Roderick M. (1966). *Theory of Knowledge*. Englewood Cliffs, N.J.,Prentice-Hall.

Choi, Sungho (2006). The simple vs. reformed conditional analysis of dispositions. *Synthese* 148 (2):369 - 379.

Choi, Sungho (2009). The conditional analysis of dispositions and the intrinsic dispositions thesis. *Philosophy and Phenomenological Research* 78 (3):568-590.

Choi, Sungho (2010). Dispositions and Bogus Counterexamples: Reply to Lee. *Philosophia* 38 (3):579-588.

Christensen, D. (2007a). Epistemology of disagreement: The good news. *The Philosophical Review*, 187-217.

Christensen, D. (2007b). Three questions about Leplin's reliabilism. *Philosophical Studies* 134 (1):43 - 50.

Comesaña, Juan (2005). Unsafe knowledge. *Synthese* 146 (3):395 - 404.

Davidson, Donald (1982). Rational animals. *Dialectica* 36 (4):317-28.

Davidson, Donald (1984). *Inquiries Into Truth And Interpretation*. Oxford University Press.

Davies, Martin (1994). The mental simulation debate. *Philosophical Issues* 5:189-218.

Dennett, Daniel C. (1983). Taking the intentional stance seriously. *Behavioral and Brain Sciences* 6 (3):379.

Dennett, Daniel C. (1985). When does the intentional stance work? *Behavioral and Brain Sciences* 8 (4):763.

Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.

Dennett, Daniel C. (1995). Do animals have beliefs? In H. Roitblat & Jean-Arcady Meyer (eds.), *Comparative Approaches to Cognitive Science*. MIT Press.

De Ridder, J. (2013). Why Only Externalists Can Be Steadfast. *Erkenntnis*, 1-15.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature.*Scientific American*, *208*, 45-53.

Dubois, D., Lang, J., & Prade, H. (1991). A brief overview of possibilistic logic. In *Symbolic and Quantitative Approaches to Uncertainty* (pp. 53-57). Springer Berlin Heidelberg.

Edgington, Dorothy (1995). On conditionals. *Mind* 104 (414):235-329

Elga, A. (2007). Reflection and disagreement. *Noûs*, *41*(3), 478-502.

Elgin, Catherine Z. (2004). True enough. *Philosophical Issues* 14 (1):113–131.

Ellis, Brian. "Marc Lange on Essentialism." *Australasian Journal of Philosophy* 83 (2005): 75 - 79.

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

Fara, Michael (2005). Dispositions and habituals. *Noûs* 39 (1):43–82.

Feldman, Richard (1985). "Reliability and Justification," *The Monist* 68: 159-174.

Feynman, R. P. (2006). *QED: The strange theory of light and matter*. Princeton University Press.

Frege, Gottlob (1997). Grundgesetze der Arithmetik, Volume I. In M. Beaney (Ed.), *The Frege Reader* (pp. 194-223). Blackwell Publishers.

Gallese, Vittorio & Goldman, Alvin (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2 (12):493-501.

Gendler, Tamar Szabó (2008a). 'Alief and belief', *Journal of Philosophy* 105, pp. 634–63.

Gendler, Tamar Szabó (2008b). 'Alief in action (and reaction)', *Mind & Language* 23,pp. 552–85.

Gendler, Tamar Szabó (2010). *Intuition, Imagination, and Philosophical Methodology*. Oxford University Press.

Gertler, Brie (2011). Self-knowledge and the transparency of belief. In Anthony Hatzimoysis (ed.), *Self-Knowledge*. Oxford University Press.

Gettier, Edmund (1963). Is justified true belief knowledge? *Analysis* 23 (6):121-123.

Gilbert, Margaret (1987). Modelling collective belief. *Synthese* 73 (1):185-204.

Gilbert, Margaret (1994). Remarks on collective belief. In Frederick F. Schmitt (Ed.), *Socializing Epistemology: The Social Dimensions of Knowledge*. Rowman and Littlefield. 235-56.

Gilbert, Margaret (2004). Collective epistemology. *Episteme* 1 (2):95--107.

Gopnik, Alison & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language* 7 (1-2):145-71.

Gopnik, Alison (1997). The scientist as child. *Philosophy of Science* 63 (4):485-514.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. The MIT Press.

Goldman, Alvin (1976). Discrimination and perceptual knowledge. *Journal of Philosophy* 73 (November):771-791.

Goldman, Alvin (1986). *Epistemology and Cognition*. Cambridge/MA & London: Harvard University Press.

Goldman, Alvin (1988). Strong and weak justification. *Philosophical perspectives*, *2*, 51-69.

Goldman, A., & Gallese, V. (2000). Reply to Schulkin. *Trends in Cognitive Sciences*, *4*(7), 255-256.

Gordon, Robert M. (1986). Folk psychology as simulation. *Mind and Language* 1 (2):158-71.

Greco, John (2000). *Putting Skeptics in Their Place*. Cambridge: Cambridge University Press.

Greco, John (2007). Worries about Pritchard's safety. *Synthese* 158 (3):299 - 302.

Greco, J. (2010). Achieving knowledge. *A Virtue-Theoretic Account of Epistemic Normativity.* Cambridge: Cambridge University Press.

Greco, John (2012). Better safe than sensitive. In K. Becker & T. Black (Eds.) *Sensitivity in Epistemology* (pp. 193-206). Cambridge: Cambridge University Press.

Gutting, G. (1982). Religious belief and religious skepticism. Notre Dame: University of Notre Dame Press.

Hájek, Alan (2003a). Conditional Probability Is the Very Guide of Life. In Kyburg Jr, E. Henry & Mariam Thalos (eds.), *Probability is the Very Guide of Life: The Philosophical Uses of Chance*. Open Court.

Hájek, Alan (2003b). What conditional probability could not be. *Synthese* 137 (3):273--323.

Hájek, Alan. (2007). The reference class problem is your problem too. *Synthese* 156 (3):563--585.

Handfield, T. (2005). Lange on essentialism, counterfactuals, and explanation. *Australasian Journal of Philosophy*, *83*(1), 81-85.

Hannon, Michael (2013). 'Knows' entails truth. *Journal of Philosophical Research* 38:349-366.

Hazlett, Allan (2010). The myth of factive verbs. *Philosophy and Phenomenological Research* 80 (3):497 - 522.

Hazlett, Allan (2012). Factive presupposition and the truth condition on knowledge. *Acta Analytica* 27 (4):461-478.

Hetherington, Stephen (1999). Knowing failably. *Journal of Philosophy* 96 (11):565-587.

Holyoak, K. J. & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128,* 3-31.

Homer (2012). *The Odyssey of Homer*. (A. Pope Trans.). Adelaide: University of Adelaide eBooks. Retrieved from http://ebooks.adelaide.edu.au/

Huemer, M. (2001). *Skepticism and the Veil of Perception*. Rowman & Littlefield.

Hunter, David (2011). Alienated Belief. *Dialectica* 65 (2):221-240.

Kelly, Thomas (2005). The epistemic significance of disagreement. In John Hawthorne & Tamar Gendler (Eds.), *Oxford Studies in Epistemology, Volume 1*. Oxford: Oxford University Press.

Kragh, H. (1990). *Dirac: a scientific biography*. Cambridge University Press.

Kuhn, Thomas S. (1996). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lange, M. (2004). A note on scientific essentialism, laws of nature, and counterfactual conditionals. *Australasian Journal of Philosophy*, *82*(2), 227-241.

Laudan, Larry (1981). A confutation of convergent realism. *Philosophy of Science* 48 (1):19-49.

Leplin, Jarrett (2007). In defense of reliabilism. *Philosophical Studies* 134 (1):31 - 42.

Lee, Jaeho (2010). Disposition, Explanation, and Causation—A Defense of the Reformed Conditional Analysis of Disposition. *Philosophia* 38 (3):569-577.

Lee, Jaeho (2011). Genuine Counterexamples to the Simple Conditional Analysis of Disposition: A Reply to Choi. *Philosophia* 39 (2):327-334.

Leitgeb, Hannes (2012a). A Probabilistic Semantics for Counterfactuals. Part A. *Review of Symbolic Logic* 5 (1):26-84.

Leitgeb, Hannes (2012b). A Probabilistic Semantics for Counterfactuals. Part B. *Review of Symbolic Logic* 5 (1):85-121.

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, David (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, David (1997). Finkish dispositions. *Philosophical Quarterly* 47 (187):143-158.

Lycan, William G. (1986). Tacit belief. In R. Bogdan (ed.), *Belief: Form, Content, and Function*. Oxford University Press.

Macquer, P. J. (1777). *Elements of the Theory and Practice of Chemistry: Trans. from the French*. Donaldson.

Maddy, P. (2007). *Second philosophy: A naturalistic method*. Oxford: Oxford University Press.

Martin, C. B. (1994). Dispositions and conditionals. *Philosophical Quarterly* 44 (174):1-8.

Mill, J. S. (2006). *On liberty and the subjugation of women*. New York: Penguin Classics.

Mitchell, Jason P. (2005). The false dichotomy between simulation and theory-theory: the argument's error. *Trends in Cognitive Sciences* 9 (8):363-364.

Moore, G. E. (1959a). A defense of common sense. In G. E. Moore (Ed.) *Philosophical Papers* (pp. 33-59). New York: Collier Books.

Moore, G. E. (1959b). Proof of an external world. In G. E. Moore (Ed.) *Philosophical Papers* (pp. 126-148). New York: Collier Books.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge/MA: Harvard University Press.

Olsson, Erik J. (2005). *Against Coherence: Truth, Probability, and Justification*. Oxford University Press.

Plantinga, Alvin (1970). World and essence. *The Philosophical Review*, *79*(4), 461-492.

Plantinga, Alvin (1974). *The Nature of Necessity*. Oxford: Clarendon Press.

Peirce, C. S. (1877). The fixation of belief. In J. Buchler (Ed.) *Philosophical Writings of Peirce* (pp. 5-22). New York, NY: Dover.

Peirce, C. S. (1878). How to make our ideas clear. In J. Buchler (Ed.) *Philosophical Writings of Peirce* (pp. 23-41). New York, NY: Dover.

Peirce, C. S. (1880). On the algebra of logic. *American Journal of Mathematics*,*3*(1), 15-57.

Plato (1997a). Meno. In John M. Cooper (Ed.) *Plato: Complete Works* (870-897). Indianapolis: Hackett.

Plato (1997b). Theaetetus. In John M. Cooper (Ed.) *Plato: Complete Works* (157-234). Indianapolis: Hackett.

Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge.

Pritchard, Duncan (2008). Sensitivity, safety, and anti-luck epistemology. In John Greco (Ed.), *The Oxford Handbook of Skepticism*. Oxford University Press.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34(4), 517-549.

Rawls, J. (1999). *A theory of justice*. Cambridge, Mass: Belknap Press of Harvard University Press.

Read, S. J., Snow, C. J., & Simon, D. (2003). Constraint satisfaction processes in social reasoning. In *Annual Meeting of the Cognitive Science Society, Boston*.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality & Social Psychology, 65*, 429-447.

Roeper, P., & Leblanc, H. (1999). Probability theory and probability semantics.*Toronto Studies in Philosophy*.

Rowbottom, Darrell P. (2007). 'In between believing' and degrees of belief. *Teorema* 26 (1):131-137.

Saxe, R. (2005a). Against simulation: The argument from error. *Trends in Cognitive Sciences* 9 (4):174-79.

Saxe, R. (2005b). Hybrid vigour: reply to Mitchell. *Trends in Cognitive Sciences* 9 (8):364-364.

Saxe, R. (2009). The neural evidence for simulation is weaker than I think you think it is. *Philosophical Studies* 144 (3):447 - 456.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and 'theory of mind'. *Mind & Language*, *14*(1), 131-153.

Schubert, Stefan (2012). Is coherence conducive to reliability? *Synthese* 187 (2):607-621.

Schulkin, J. (2000). Theory of mind and mirroring neurons. *Trends in cognitive sciences*, *4*(7), 252-254.

Schwitzgebel, Eric (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91 (4):531-553.

Shogenji, Tomoji (2005). Justification by coherence from scratch. *Philosophical Studies* 125 (3):305 - 325.

Shogenji, Tomoji (2007). Why does coherence appear truth-conducive? *Synthese* 157 (3):361 - 372.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1250-1260.

Simpson, R. M. (2013). Epistemic peerhood and the epistemology of disagreement. *Philosophical studies*, *164*(2), 561-577.

Sommers, F. (2009). Dissonant beliefs. *Analysis*, *69*(2), 267-274.

Sosa, Ernest (1996). "Postscript to 'Proper functionalism and virtue epistemology'," in J. L. Kvanvig (ed.), *Warrant in Contemporary Epistemology*, Lanham, MD: Rowman & Littlefield.

Sosa, Ernest (2000). "Skepticism and contextualism," *Philosophical Issues*, 10: 1–18.

Sosa, Ernest (2009). *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. OUP Oxford.

Sosa, Ernest (2011). *Reflective Knowledge: Apt Belief and Reflective Knowledge, Volume II*. OUP Oxford.

Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality & Social Psychology, 62*, 913-933.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues*, *49*, 147-165.

Stalnaker, Robert (1976). Possible worlds. *Noûs*, *10*(1), 65-75.

Stalnaker, Robert (1980). A theory of conditionals. In *Ifs* (pp. 41-55). Springer Netherlands.

Stalnaker, Robert (2003). On what possible worlds could not be. In R. Stalnaker (Ed.) *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays* (pp. 40 – 54). Oxford: Clarendon Press.

Strevens, Michael (2004). The causal and unification approaches to explanation unified—causally. *Noûs* 38 (1):154–176.

Strevens, Michael (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Thagard, Paul. (1992). *Conceptual Revolutions*. Princeton University Press.

Thagard, Paul (2002). *Coherence in Thought and Action*. The MIT Press.

Thagard, Paul (2006). *Hot Thought: Mechanisms and Applications of Emotional Cognition*. Cambridge MA: Bradford Book/MIT Press.

Tsohatzidis, Savas L. (2012). How to forget that 'know' is factive. *Acta Analytica* 27 (4):449-459.

Turri, John (2011). Mythology of the factive. *Logos and Episteme* 2 (1):143-152.

Van Fraassen, B. C. (1966). Singular terms, truth-value gaps, and free logic. *The journal of Philosophy*, 481-495.

Van Fraassen, B. C. (1968). Presupposition, implication, and self-reference. *The Journal of Philosophy*, 136-152.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P. & Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage*, *14*(1), 170-181.

Wray, Brad K (2001). Collective belief and acceptance. *Synthese* 129 (3):319-33.

Zimmerman, Aaron Z. (2007). The nature of belief. *Journal of Consciousness Studies* 14 (11):61-82.

APPENDIX A: A SUPERVALUATIONAL SEMANTICS FOR THE COUNTERFACTUAL CONDITIONAL

The primitives in our counterfactual language L are as follows:

i. A countably infinite set of *individual variables*
ii. For every n, a countably infinite set of n-ary *predicates*
iii. The *connectives* $\supset$ (material implication) and $\neg$ (negation)
iv. The *counterfactual conditional connective*, $>$
v. The *universal quantifier*, $\forall$
vi. The *identity sign*, $=$
vii. The *auxiliary symbols* for grouping, (and)

Our system is autonomously named. As such, x, y, and z will be used as metavariables for (i); P^n will be used as a metavariable for (ii), with the superscript possibly omitted in instances where doing so does not engender ambiguity. I will also use the common definitions of & and $\vee$ from (iii), and the common definition of $\equiv$ from & and $\supset$. Similarly, the common definition of $\exists$ from (v) will be used. Any string of (i) - (vii) is a *formula*. A *formula* is a *well-formed formula* (or *wff*) by the following definition:

i. A formula of the form $P^n x_1 \ldots x_n$ is a wff
ii. A formula of the form $x = y$ is a wff
iii. If A is a wff, then $\neg A$ is a wff
iv. If A is a wff and x is a variable, then $\forall xA$ is a wff
v. If A and B are wffs, then $A \supset B$ is a wff
vi. If A and B are wffs, then $A > B$ is a wff
vii. Nothing else is a wff

I will use the metavariables A, B, and C as metavariables for wffs. The notions of a variable being *free* or *bound* are the usual, as is the notion of a *well-formed part* (wf part) of a wff. Similarly, A^y/x and $A^y//x$ receive their common definitions, and may be read as "y replaces all x in A" and "y replaces at least 0 occurrences of x in A" respectively.

The notion of a model-structure for L is as follows: the *domain, D*, is a non-empty set. The *interpretation function*, f, is a unary function total on the set of individual variables of L and

total on the set of predicates of *L,* which assigns each individual variable of *L* an element of *D*, and assigns to every *n*-ary predicate of *L* a set of ordered *n*-tuples of elements of *D*. A *model structure*, then, is an ordered pair $M = \langle D, f \rangle$. The more difficult part of my semantics, the notion of a wff's truth in a model structure, is two-tiered. First, I will define the *primary auxiliary valuation* which takes care of those matters of *L* which do not involve the counterfactual conditional > having a false antecedent. I will then define the *secondary auxiliary valuation,* which will make precise the notion of counterfactual truth and will account for wffs containing >. The *valuation* will combine the two.

The primary auxiliary valuation V^*_M relative to a model structure *M* is the partial unary function *G* from the set of all wffs of *L* to {⊤, ⊥} such that:

i. If *A* is of the form $P^n x_1 \ldots x_n$: if $\{ f(x_1), \ldots, f(x_n) \} \in f(P)$ then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
ii. If *A* is of the form $x = y$: if $f(x) = f(y)$ then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
iii. If *A* is of the form $\neg B$: if $G(B)$ is ⊥ then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
iv. If *A* is of the form $\forall x B$: if $G(B^y/x)$ is ⊤ for every individual variable *y* then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
v. If *A* is of the form $B \supset C$: if $G(B)$ is ⊥ or $G(C)$ is ⊤ then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
vi. If *A* is of the form $B > C$ and $G(B)$ is ⊤: if $G(C)$ is ⊤ then $G(A)$ is ⊤, and otherwise $G(A)$ is ⊥.
vii. $G(A)$ is not defined if not by virtue of (i)-(vi).

The above establishes what "standard" truth is, and accounts for uses of the counterfactual conditional which are not counter to fact. To establish the truth of a counterfactual sentence containing a false antecedent, we perform a restriction-completion procedure on our model structure *M*. We first restrict the model such that it is compatible with our antecedent; we then

perform completions of this model and supervaluate over them. I understand completion, restriction, and the like in the following way.

A model structure $M' = <D', f'>$ is a *restriction* of a model structure $M = <D, f>$ just in case the following hold: (1) $D' \subseteq D$, (2) For every individual variable x of L defined in M', $f'(x) = f(x)$, (3) For every predicate P of L, $f'(P) \subseteq f(P)$. A model structure $M' = <D', f'>$ is an *expansion* of a model structure $M = <D, f>$ just in case M is a restriction of M'.[1]

A well-formed formula that does not contain the counterfactual conditional, A, is *compatible* with a model structure M just in case an expansion M' of M satisfies A. A model structure M' is an *A-restriction* of M just in case M' is a restriction of M, M' is compatible with A, and M' is compatible with ~A. M' is a *minimal A-restriction* of M just in case M' is an A-restriction of M and there is no A-restriction of M which is an expansion of M'.

A model structure M' is a *minimal A-expansion* of M just in case M is a minimal A-restriction of M' and M' satisfies A.

It must be noted that restrictions and expansions of a model structure $M = <D, f>$ may have interpretation functions partial on the set of individual variables of L. If M's interpretation function is total on the set of individual variables of L, then it is said to be *complete*, and is a *completion* of every model structure $M' = <D', f'>$ such that M' is a restriction of M.

If a well-formed formula S does not contain the counterfactual conditional >, then it is to be valuated in keeping with standard semantics: S is true just in case M satisfies S; if M does not satisfy S, then S is false. If S is of the form $A > C$ and M satisfies A, then S is true just in case M

---

[1] By requiring that f is a *total* function on the individual variables and predicates of L, I am assuming that every singular term of L denotes. As such, sentences like "If Pegasus had wings, then something identical to Pegasus would exist" cannot be adequately expressed in L.

satisfies C, and false otherwise. If M does not satisfy A, we generate a minimal A-restriction of M to obtain M'. If we are unable to complete this step (for example, if A is of the form p & ~p), then S receives no truth-value. Having generated M', we next construct a minimal A-expansion of M' to obtain $M\dagger$. $M\dagger$, as a minimal A-expansion of a minimal A-restriction, amounts to a "pruning" of our original model to allow it to satisfy A, and the follow-through of altering it to satisfy A. In our new model $M\dagger$, then, A is the case. If $M\dagger$ satisfies C, then S is true. If $M\dagger$ does not satisfy S, we perform a supervaluation over all completions of $M\dagger$.[2] If all completions satisfy C, then S is true. If all completions do not satisfy C, then S is false. If some completions satisfy C while others do not, then the truth-value of S is undefined.[3]

I turn now to the formalization of my secondary auxiliary valuation, which accounts for counterfactual truth in a model $M\dagger$ which is a completion of a minimal A-expansion of M. The secondary auxiliary valuation $V^*_{M\dagger}{}^*_{M'}{}^*_{M}$ relative to a model M, a minimal A-expansion of M M', and a completion of a minimal A-expansion of M $M\dagger$ is the total unary function G from the set of all wffs of L to $\{\top, \bot\}$ such that:

> If $M\dagger = \emptyset$, then $G(A)$ is undefined.
> If A is an atomic wff and $V^*_{M}(A)$ is defined, then $G(A) = V^*_{M}(A)$.
> If A is an atomic wff and $V^*_{M}(A)$ is not defined, then $G(A) = V^*_{M\dagger}(A)$.

Because only wffs A of the form $B > C$ can be undefined after an application of the primary auxiliary function, and $M\dagger$ is the final step of the restriction-completion process in which B

---

[2] Note that if $M\dagger$ satisfies C, then all completions of $M\dagger$ will satisfy C as well. Thus our examination of $M\dagger$ as to whether C seems redundant. However, because we are working with the notion of a mental construction, we perform this relatively simple task as a labor-saving step, so that we do not go through the process of constructing completions of $M\dagger$ when such mental experiments will fail to give us new information regarding C.

[3] That is to say, S has no truth-value. An undefined truth value must not be mistaken for a truth value of "undefined."

holds, A will receive a truth value when B is false by step (ii) from $V^*_{M\dagger}{}^*_M$ function combined with step (vi) of $V^*_{M\dagger}$ (which is, of course, the primary auxiliary function as applied to $M\dagger$).

Our final *valuation* V_M relative to a model M is the supervaluation over all the secondary valuations $V^*_{M\dagger}{}^*_{M'}{}^*_M$ where $M\dagger$ is a completion of a minimal A-expansion (M') of M. It is the partial unary function G from the set of all wffs of L to $\{\top, \bot\}$ such that:

> If $G'(A) = \top$ for every secondary auxiliary valuation G' relative to M, a minimal A-expansion of M M', a completion of M' $M\dagger$, then $G(A) = \top$.
> If $G'(A) = \bot$ for every secondary auxiliary valuation G' relative to M, a minimal A-expansion of M M', a completion of M' $M\dagger$, then $G(A) = \bot$.
> $G(A)$ is not defined if not by virtue of (i)-(ii).

APPENDIX B: ECHO SCRIPTS FOR COHERENTIST MODELS

What follows are models in ECHO-script. Where possible, results were verified with an exhaustive algorithm as well as a connectionist algorithm. Both Fallibility (Section 1) and Infallibility (Section 2) were correctly solved by the connectionist algorithm; The Mess (Section 3) presents a case that required an exhaustive algorithm. The ECHO script was generated by ConvinceMe-mod, the author's modification of Patricia Schank's original implementation of Thagard's ECHO algorithm. ConvinceMe-mod allows for joint contradiction (i.e. two propositions *p* and *q* exclude a third *r*, but each of *p* and *q* is individually compatible with *r*), competition, and simplicity-preference and implements an exhaustive algorithm that, though computationally quite expensive, is guaranteed to provide an optimum solution in those instances in which its computation is practically feasible.

1. Fallibility

This model accepts H9, that at least one of my beliefs H1-H7 is wrong, while also individually accepting H1-H7. In light of past error, it rejects the proposition that I am infallible, i.e. that *all* of the accepted propositions H1-H7 are true. Connectionist and exhaustive algorithms agree on the solution.

```
data (E1, E2, E3, E4)
explain ((H1, H2, H3, H4, H5, H6, H7), E1)
explain ((H1, H2, H3, H4, H5, H6, H7), E2)
explain ((H9), E3)
explain ((H9), E4)
contradict (H8, H9)
contradict (H8, E3)
//NB: This is a joint contradiction and will not work with an
//unmodified ECHO interpreter!
jcontradict((H1, H2, H3, H4, H5, H6, H7), H9)
contradict (H8, E4)
```

2. Infallibility

This model rejects H3 ("H1 or H2 is false") while accepting both H1 ("I exist") and H2 ("H1 is true"). Connectionist and exhaustive algorithms agree on the solution.

```
data (E1)
explain ((H1), H2)
explain ((H1), E1)
//NB: This is a joint contradiction and will not work with an
//unmodified ECHO interpreter!
jcontradict((H1, H2),H3)
```

3. Mess

In this model, connectionist and exhaustive algorithms disagree. The optimum solutions found by the exhaustive algorithm are acceptance of H1 and H3 and rejection of H2 and H4 or *vice versa*. Each case has a coherence measure of 0.60, and all dispositional links among propositions are satisfied. The connectionist algorithm accepts all propositions, yielding a coherence measure of -0.36. This model is included to show that although the connectionist algorithm is often useful, it does have its limitations. In such circumstances, an exhaustive algorithm is much preferable insofar as it is computationally feasible.

```
data (E1, E2, E3, E4)
explain ((H1), E1)
explain ((H2), E2)
explain ((H3), E3)
explain ((H4), E4)
contradict (H1, H2)
contradict (H2, H3)
contradict (H3, H4)
contradict (H4, H1)
```

APPENDIX C: THE CALCUHATER

```
m=-1
n=-1

#sanity-checked input
while (m < 1000 || m > 10000)
        puts "Enter a number between 1,000 and 10,000:"
        m = gets.strip.to_i
end
while (n < 1000 || n > 10000)
        puts "Enter a number between 1,000 and 10,000:"
        n = gets.strip.to_i
end
sum = m + n

#Picks a number 1-10, subtracts 5 yielding a number from -5..5,
#then multiplies by 10 so that the last digit is correct.
hate = (rand(11) - 5) * 10
result = sum + hate
puts "The sum is #{result}."
```