CAUGHT IN THE WEB:
INTERNALIZING A NATURALISTIC THEORY OF EPISTEMIC JUSTIFICATION

By

MATTHEW R. PIKE

B.A., University of Colorado Denver, 2001

M.A., University of Colorado Boulder, 2007

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Philosophy

2017

This thesis entitled:
Caught in the Web: Internalizing a Naturalistic Theory of Epistemic Justification
written by Matthew R. Pike
has been approved for the Department of Philosophy

_____
Robert Rupert, committee chair

_____
Michael Tooley, committee member

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Pike, Matthew (Ph.D., Philosophy)

Caught in the Web: Internalizing a Naturalistic Theory of Epistemic Justification

Thesis directed by Professor Robert Rupert

Most proponents of a naturalistic approach to epistemology seem to feel forced to endorse a process reliabilist theory of justification, ostensibly to forestall charges that their naturalistic views can yield only a descriptive account of belief that is devoid of normative force. This reliabilist approach to justification depends upon an externalist grounding, holding that belief-fixing and sustaining processes and procedures are reliable because they generally produce beliefs that are objectively *true*. This dissertation will explore some of the problems for the standard externalist approach that naturalists usually favor, and then show that, while the frequent conjoining of naturalist epistemology and process reliabilism found in the literature might lead one to believe that there are strong implications from epistemic naturalism to externalism, there is in fact a largely unexplored philosophical space that combines naturalized epistemology and justificatory internalism. A version of process reliabilism that is compatible with both naturalism and internalism will then be developed and defended from several potential objections. This theory of justification will demonstrate that internalism concerning epistemic justification is more compatible with naturalistic epistemological commitments than has previously been noted, and that many naturalists, perhaps having failed to fully consider this option, may have been too hasty in adopting externalist views of justification.

For my beloved family, friends, and mentors—

especially those that are all three.

You provide meaning in this reality,

whatever its nature may be...

Contents

Figures

Chapter 1: Introduction

It's 9:00am on a typical day. People have dressed, had their morning coffee, and are heading out to accomplish their various tasks for the day. Business people are arriving at their places of work, intent on making the day profitable; students are headed to classes to learn new things (or, at least, to figure out how to pass their courses); teachers are preparing lectures in their heads, and making copies of assignments to evaluate their students' performance; scientists are headed into their labs to test new hypotheses; and philosophers are settling into their arm chairs to try to discover deep and interesting truths. All of this is made possible by the astonishingly complex brain that each human possesses. But just what are these brains doing? According to most epistemologists, all of these people have in common the tacit goal of arriving at true beliefs by adhering to justificational norms, and the varying degrees of success their brains have in attaining this goal will (at least largely) account for their varying degrees of success in accomplishing their individual goals for the day. The "more rational" people will be more likely to survive crossing the street, get promoted in their jobs, pass their course exams, and get their research findings published because their cognizing better follows and conforms to the correct set of epistemic norms. Individuals who follow these norms are rewarded with the attainment of justified beliefs, and since these beliefs are justified, many of them are also, hopefully, true. But exactly what is this justification that plays such a key role in all of our activities?

1.1 Overview

Epistemology has had justification as one of its key targets since at least the time of Plato's *Theaetetus*. While these philosophical investigations have produced interesting and informative theories, their practitioners did not have much of the scientific information available that we do today. Recent developments in neuroscience, evolutionary psychology, evolutionary biology, cognitive ethology, and other fields, as well as the extensive evidence for the success and fruitfulness of scientific inquiry, both afford and demand the reworking of our epistemic theories. However, a worrisome gap has developed between these formal fields of scientific inquiry and philosophical epistemic inquiry, as many philosophers continue to attempt to modify and improve the traditional epistemic theories, which ultimately date back to the likes of Plato, Aristotle, and Descartes.

Contemporary epistemology has continued to place a heavy emphasis on investigating the nature of *justification* (which is closely related to epistemic *warrant* that turns true belief into *knowledge*).[1] While this has always been an important component of epistemology, ever since Gettier (1963) demonstrated the inadequacy of the classic "justified true belief" theory of knowledge, there has been an almost desperate interest by many epistemologists in improving and reformulating theories of justification. This may be driven by a desire to shore up our theories of knowledge from Gettier's worries, and since a traditional, but problematic, notion of justification is frequently diagnosed as being what allows "Gettier problems" to occur, an effort to salvage a theory of knowledge requires new theories of justification. And indeed, there has

[1] Pollock and Cruz think that it is an unfortunate feature of contemporary epistemology that so much research is focused on what must be added to the "Justified True Belief" theory of knowledge to solve the Gettier problem, and that analyzing justification has *always* been the central project that epistemology should address (1999, p. 13-14).

been an explosion of new theories of justification in the literature aimed at solving Gettier-style problems. Some epistemologists,[2] however, think that Gettier's results are even more damaging, and force us completely to abandon knowledge as a useful topic of inquiry. Instead, it is claimed, epistemology should focus on justified belief and epistemic norms, since these notions both capture what we want epistemology to do *and* might still be feasible in a post-Gettier environment. Either way, justification is one of the central concerns for practicing epistemologists today, and this is reflected in the amount of research produced on the topic.

At the same time, there has also been an important trend toward a naturalist conception of epistemology. Impressed by the enormous success of science, and in particular the progress made in the various fields of psychology and neuroscience, many epistemologists have sought to reorient epistemology in a manner more friendly to, and influenced by, the natural sciences.[3] This approach marks a commitment to abandoning the "spooky" or "super-natural" components that are seen by naturalists as having made their ways into traditional philosophical accounts.[4]

---

[2] Including, for example, John Pollock and Joseph Cruz (Pollock and Cruz, 1999)

[3] In the discussion that follows, it will be helpful to keep in mind that the *success* of science can be understood in at least two importantly different ways. Scientific realists would generally claim that science's success has been at discovering either the truth (or at least "approximate truth"—see Psillos [1999] for discussion) about some of the entities that make up the existent world and their properties, or in developing scientific models that capture and describe structures and relations that really exist in the world (see Da Costa and French [2003] for development of a notion of "partial truth" to use in evaluating models and their isomorphism to target systems). Scientific anti-realists, such as Bas van Fraassen, while still agreeing that science has been enormously successful, would say instead that the success of science is to be found in the construction of *empirically adequate* models (or *structures*). These models can still be enormously useful in helping us to attain our goals, whether or not the structures and relations modeled are present in the actual world. As will become apparent, I am suspicious of the realists' claims, and find the pragmatic victories of science to be more than sufficient to motivate our interest in employing scientific practice more widely, and so it is in this pragmatic, anti-realist sense that I mean that science has shown itself to be successful.

[4] Goldman (1979/1994), Kornblith (2002), and P.S. Churchland (1987) are just a few examples that express this commitment.

At the intersection of these two trends, interesting questions arise when one investigates how this focus on justification has manifested within the naturalist approach to epistemology. Of particular interest is the popular theory of justification known as *process reliabilism*, and how it is situated within naturalized epistemology. Many proponents of a naturalistic approach to epistemology seem to feel forced to endorse a process reliabilist notion of justification, or a modified version of it,[5] ostensibly to forestall charges that their naturalistic views can yield only a descriptive account of belief that is devoid of normative force. This reliabilist approach to justification depends upon an externalist grounding,[6] holding that belief-fixing and sustaining processes and procedures are reliable because they generally produce beliefs that are objectively *true*. Justificatory externalism (roughly, the view that *at least some* factors involved in the justification of an epistemic agent's beliefs are external to the agent), however, is controversial in the epistemological literature, and well-discussed potential problems with it indicate that internalism (the alternative view that the justification of a belief results *only* from internal factors such as its relation to other beliefs, memories, perceptions, and so forth), is still very much a live philosophical option.[7] Given this, we might wonder why naturalism has traditionally been so

[5] Some examples include: Goldman (1979-2012), Grundmann (2009), Antony (2004), Comesaña (2002), Henderson and Horgan (2007), and Lyons (2009). Churchland (2007, ch.6) also advocates for an interesting version of process reliabilism, one that is completely "liberated from propositional attitudes". There are also numerous theories that have been heavily influenced by process reliabilism. For example, Bishop and Trout (2005) develop "Strategic Reliabilism" an approach that focuses on cognitively efficient problem solving using reliable strategies and statistical prediction rules.

[6] Comesaña states that, "reliabilism is marketed as a version of externalism—indeed, as the *paradigmatic* externalist theory" (2010, p.577, original emphasis).

[7] See Kornblith, 2001 for extensive discussion of the ongoing internalism versus externalism debate. Even the correct way to delineate internalism from externalism is highly controversial. For example, on the contemporary view of internalism known as *accessibilism*, the justificatory status of a belief completely depends on relata that must be accessible to the epistemic agent upon reflection (see, for example, BonJour, 1985, p. 16-33 and Chisholm, 1988). However, a popular alternative view, *mentalism*, adheres

closely connected to externalist theories of justification, and whether this connection is necessary. As Pollock and Cruz have observed, Alvin Goldman is both the best-known naturalist epistemologist and the best-known epistemic externalist, and the mere familiarity of his work may have played an important role in the association of the two approaches.[8]

While this frequent conjoining of naturalist epistemology and process reliabilism found in the literature might lead one to believe that there are strong implications from epistemic naturalism to externalism, there is a largely unexplored philosophical space that combines naturalized epistemology and justificatory internalism. This dissertation explores and motivates an internally-driven, coherence-based version of process reliabilism which is defined endogenously and uses coherence measures (including some formal tools proposed by Bovens and Hartmann, 2003) to provide a theory that allows us to differentiate between proper and improper belief formation, while maintaining a purely naturalistic, internalist account of our epistemic processes. This demonstrates that internalism concerning epistemic justification is more compatible with the holding of naturalistic epistemological commitments than has previously been noted, and that many naturalists, perhaps having failed to fully consider this option, may have been too hasty in adopting externalist views of justification.

This dissertation consists of six chapters. Chapter 1 continues with an introduction to the two basic commitments of naturalism: the metaphysical (or ontological) component and the closely related methodological one. Once the central commitments of a naturalist approach to

---

to the weaker requirement that the relata be part of one's mental life, sometimes reflectively accessible, but sometimes not (Conee and Feldman, 2004, p. 53-82). Both of these approaches to internalism also face questions about whether the beliefs or mental states that are relevant must be occurrent or not. At this point, however, I want to remain neutral on how best to understand internalism, as later in the dissertation I argue for a new line of demarcation between internalism and externalism.

[8] Pollock and Cruz, 2004, p. 139

epistemology have been delineated, the mainstay naturalist epistemic theory of justification, process reliabilism, is presented and described in detail before the chapter finishes with an identification of what I take to be the most important components of the standard process reliabilist approach.

Chapter 2 discusses multiple objections to process reliabilism to demonstrate why a shift to a different naturalist theory of justification is necessary. The first section considers the most commonly discussed objections to externalist process reliabilism found in the philosophical literature, including the difficulty of delineating in which environments (or possible worlds) a process must be reliable in order to confer justification, the Cartesian evil-demon counter-example, BonJour's "reliable clairvoyance" counter-example (which seems to give an example of an individual with reliable cognitive processes who ought not be seen as being justified in holding the resulting beliefs), and the generality problem (which asks how the "cognitive processes" that are key components in a process reliabilist account of justification are to be differentiated). The rest of the chapter raises additional, less commonly discussed problems for process reliabilism.

Chapter 3 draws upon the lessons learned from the problems discussed in Chapter 2 to construct a series of *desiderata* that a theory of justification should meet. Specifically, this chapter examines the motivation for preferring an internalist, rather than an externalist, theory of justification; argues for an end to restricting epistemic theorizing to doxastic states (such as beliefs) and for instead including subdoxastic states as central to any correct theory of justification; and argues that an improved theory of justification should be pragmatic in nature. These goals and constraints that result from studying the defects of standard process reliabilism then provide the framework for the theory of justification that I offer in the next chapter.

Chapter 4 presents the theory of justification, termed "endo-reliabilism", which aims to incorporate the strengths of standard process reliabilism while also meeting the *desiderata* identified in Chapter 3. The development of endo-reliabilism begins with a discussion of one very promising approach to precisely modeling reliability,[9] which provides the framework and tools necessary to develop the new theory on offer. The theory is then elaborated, and some examples of its application to real-world cases are considered.

Chapter 5 defends the theory developed in Chapter 4 by responding to several possible objections, including a worry that an internalist framework is incompatible with naturalism, an objection that the computational power required by the theory on offer is too heavy to be realized in brains like ours, and an objection that the theory as developed is not adequately *normative*, and so fails as a theory of justification. These objections, while perhaps initially tempting, ultimately are shown to pose no real threat to the endo-reliabilist theory of justification.

Chapter 6 then discusses some additional ramifications of the new theory developed in Chapter 4, and suggests some broader applications of the theory, as well as directions for further research.

1.2 Naturalism: Commitments and Methods

While evaluating all of the work done advocating a naturalistic approach is outside the scope of what can be addressed here, it will be beneficial to explicate some of the central commitments of naturalist epistemology, and say just a little about their history.

Some of the central ideas of naturalistic epistemology can be found in the work of earlier philosophers, but current versions of naturalist epistemology can almost universally be traced

[9] Bovens and Hartmann (2003)

back to Quine's paper "Epistemology Naturalized" (Quine, 1969). The further development of this family of views in the literature, however, has branched considerably. While naturalism now comes in many forms, it is typically seen as involving two primary commitments. The first is *ontological naturalism*, which holds that only the types (or kinds), objects, and properties that are found in and utilized by our best scientific theories exist. Examples of these entities include atoms, ionic bonds, electrical charge, and so on.

This ontological commitment entails one of the fundamental projects of naturalistic epistemology, which is to translate or "reconceive" traditional, value-laden epistemic concepts like knowledge, justification, having "good reasons" and so forth into the language of such sciences as psychology and neuroscience.[10] [11] In "What is Justified Belief?" (1979/1994) Alvin Goldman gives a list of examples of epistemically evaluative terms that ought to be avoided in our epistemic theorizing. He thinks that when we are trying to ascertain what justification is, terms like the following must be avoided if our theory is to satisfy the naturalist's ontological leanings, and be properly illuminating and explanatory: "justified, warranted, has (good) grounds, has reason (to believe), knows that, sees that, apprehends that, is probable (in an epistemic or inductive sense), shows that, establishes that, ascertains that" (Goldman, 1979/1994, p. 106). Indeed, these terms do not seem to refer to the kinds of things that we expect to encounter in any of the natural sciences, and since they are themselves epistemically evaluative,

[10] This is closely related to what Steup calls *Analytic Naturalism* where the "epistemological task is to specify in nonnormative terms on which nonnormative properties epistemic justification supervenes" (1996, p.191).

[11] This project makes contact with numerous questions within the philosophy of mind, such as whether our so-called "folk psychology", which covers human mental states like beliefs and desires, should be seen as reducible to our best scientific theories of the mind, or should be eliminated in favor of them. See Churchland (1981, 1989) and Dennett (1971, 1981, and 1991) for examples of the discussion about this issue.

they do not shed any explanatory light on the epistemic notions they are often used to discuss. These terms may be reducible to things like brain states, and relations, but if that is the case, a naturalized epistemology should talk about epistemic notions in *those* terms instead. Examples of other terms that Goldman thinks *are* permissible are: "believes that, is true, causes, it is necessary that, implies, is deducible from, is probable (either in the frequency sense or the propensity sense)" (Goldman, 1979/1994, p. 106). These terms are not epistemic in nature, and so can be used in a theory that attempts to give a naturalist grounding to our epistemic notion of justification.[12]

Exclusively accepting the ontology of science has an important ramification that should be explicitly acknowledged before proceeding further. Since epistemological questions concern things like what an "epistemic agent" or a "subject with a mind" ought to believe, the background assumptions made about the nature of the mind will play an important role in shaping the answers that are reached. As is standard in naturalist epistemological work, I will assume for the remainder of what follows that human minds (as well as the minds of any other animals on our planet that have them) are purely physical. Our best scientific investigations into the nature and workings of the mind have continually approached it as a physical thing, forming hypotheses and conducting experiments oriented around observable behaviors, gross brain architecture and fine neural details at the level of neurons, synapses, neurotransmitters and so on, and these approaches have indeed been incredibly successful. Once a philosopher accepts

[12] Of course, some of these terms are controversial. Paul Churchland, for example does not accept the *folk psychological* term 'believes that' as properly naturalized, since he think it is purely a left-over term from an outdated and false theory of psychology. If he is right, then a correct and fully naturalized theory of epistemology must instead make use only of the types of things that a correct and complete psychology or neuroscience discovers. See his "Eliminative Materialism and the Propositional Attitudes" (1981) for discussion.

ontological naturalism, it seems there are only three types of theory about the nature of the mind from which to choose. The first possibility, non-reductive physicalism, is not a good option for a committed naturalist. This view holds that (at least some) mental states *supervene* on the physical material, organization, and activity of our bodies; and have causal powers of their own; but are not reducible to the realm of neurons and other purely physical matter and the accompanying physical properties. What, exactly, reduction requires is controversial, but Jaegwon Kim describes the basic relationship as follows,"[i]f Xs are reduced, or reducible, to Ys, there are no Xs *over and above* Ys—to put it another way, there are no Xs *in addition to* Ys" (Kim, 2006, p. 275, original emphasis).[13] For example, modern science has successfully reduced heat to the mean kinetic motion of molecules; once the proper story is told about happenings at the molecular level there is not an additional thing, heat. So if non-reductive physicalism is correct, and some mental states are not reducible to physical matter and properties, then even a completed neuroscience that has learned everything about the physical entities and properties of the brain will be incomplete and still will need some mental states *added* before a complete story is available.

The theory of non-reductive physicalism, however, has been shown to be unstable by Kim (1993). The basic problem that Kim identifies is that, since physicalists should readily accept a principle of causal closure (which holds that any physical effect has a physical cause that was sufficient to bring about the effect in question), non-reducible mental states *cannot have any causal powers* without resulting in causal overdetermination. This means that one must either reject causal closure, reject the causal efficacy of mental states, or accept presumably

[13] Kim (2006, especially p. 273-288) provides an informative and influential analysis of reduction and the attendant complications.

widespread overdetermination. None of these options seems promising, and so non-reductive physicalism should be, and typically is, rejected by the naturalist.

This leaves only two viable theories of mind from which the ontological naturalist may choose. Either mental states, properties, types, and experiences *are* completely reducible to the physical "stuff" that our natural science studies (in which case we accept some version of reductive-physicalism) or some of them are not reducible and so must be eliminated as non-scientific, non-real, and, well, just plain "spooky" (in which case we are eliminativists about those things). While it is an interesting and important question which of these is the correct philosophical route to take, at this point I merely wish to note that one of these two views will be assumed to be correct for the remainder of the discussion.[14]

Let us turn now to the second typical commitment of naturalism, which takes the form of *methodological naturalism*. This is based on the view that philosophy and science are not actually as different as most philosophers have traditionally assumed, since philosophy and science in fact have a similar aim— which could perhaps roughly be described as "investigating the way that things are". As a result of this similarity, naturalists generally hold that philosophy should be done utilizing the same methodology employed by the natural sciences, which seems obviously to have been enormously successful in its applications to date. (One need only look briefly at our medical capabilities, wireless technology, and so forth to see clearly how far our species' recent focus on scientific methodology has taken us.) Science seems to be more adept at providing us useful insights into the nature of the world than more traditional, non-scientific methods. As Goldman says

[14] It is unfortunately not possible to engage the extensive philosophical literature on this topic here, but for a few of what I consider the stronger arguments in support of physicalist or eliminativist approaches to mind, see Dennett (1991), and Churchland (1981, 1989, 1995, 2007, and 2012).

> [o]ur folk understanding… has a limited and tenuous grasp of the processes available to the cognitive agent. Thus, one important respect in which epistemic folkways should be transcended is by incorporating a more detailed and empirically based depiction of psychological mechanisms. Here too epistemology would seek assistance from cognitive science. (Goldman, 1993, p. 273)

This science-driven approach advocated by epistemic naturalists suggests that epistemological questions should be investigated using experiments in scientific fields (such as neuroscience, cognitive science, computer science, evolutionary psychology, cognitive ethology, and so on), and that ultimately, our epistemological theories should also make heavy, if not exclusive, use of the terms and objects found in these sciences (like differing cognitive architectures, neural activations, neural signal inhibition, genetic information transmission, etc.), and the empirical experiments investigating them and their properties.[15]

Quine was so committed to this approach that, in some places, he even (notoriously) advocated the replacement of the philosophical project of epistemology by psychology (Quine, 1969/1994).[16] While I do not advocate anything quite as extreme as Quine's "replacement

[15] These two commitments of naturalism are closely connected. On one hand, it could be argued that the methodology commitment is largely responsible for the ontological views of naturalists. It is from applications of the scientific method to experiments in particle physics that scientists have come to accept entities like electrons and quarks, and are coming to accept Higgs Bosons particles as *real* things. And it is because of science that we now reject things like witchcraft and demonic possessions as unreal (well, *most* scientists reject these things, anyway). One of the central aims of science is to answer questions about what exists and what does not and to settle eventually upon the correct ontology.

On the other hand, the *natural* sciences exist specifically to discover the causal interactions and other relationships that hold between the things that we find in the world around us, and what drove and continues to shape the development of scientific methodology is finding what "works" with the entities with which we are concerned. By starting with an ontology (open to modification as needed, of course), we are able to frame questions about the "things" in question and their interrelations so that we can then bring the astonishingly successful scientific methodology to bear on this topic of inquiry.

[16] The rest of Quine's work shows that he was not really committed to this broad and drastic of a suggestion, but the idea has exerted considerable influence in the discussion of naturalism. Criticisms of his claim here often ignore the fact that it was targeted at traditional, "arm chair", Cartesian-style foundationalism that aims at certainty. Quine's other work suggests that he is more amenable to the continuation of a suitably modified and improved epistemological project, and finds it important to retain a normative element. For example, Quine later writes, "Naturalization of epistemology does not jettison the normative and settle for the indiscriminate description of ongoing procedures. For me normative epistemology is a branch of engineering. It is the technology of truth-seeking, or, in a more cautiously epistemological term, prediction.... There is no question here of ultimate value, as in morals; it is a matter

thesis", I do think that it is clearly important for our epistemic theories to "take science seriously" and derive epistemic accounts that are both compatible with, and amenable to investigation by, scientific inquiry. The theory of justification that I offer in this dissertation aims to meet this requirement, but before turning our attention to it, let us briefly examine the main "popular" naturalist theory of justification that resulted from previous attempts to satisfy the same ontological and methodological commitments: process reliabilism.

1.3 Process Reliabilism as the Naturalist's Standard

The best-known version of process reliabilism (as a theory of justification) is the one developed by Alvin Goldman (1979/1994), and modified multiple times by him in subsequent publications. Since many naturalist epistemologists hold a theory that is very similar to Goldman's views (especially those found in his earlier work from 1979 and 1986),[17] it should be sufficient for the discussion at hand to focus on the central elements of this widely discussed version of reliabilism.

When we ask whether a belief is justified or not, on one standard, traditional interpretation what we are asking is whether the belief *ought* to be accepted as true (or at least whether it is epistemically *permissible* for the agent to accept it), without directly knowing whether the particular belief in question actually *is* true or not.[18] It seems then, that most

---

of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed." (Quine, 1986, p. 664)

[17] See, for example, Grundmann (2009), Lyons (2009), Henderson and Horgan (2007).

[18] Later sections will evaluate whether this kind of interpretation can be accommodated by a naturalistic theory.

epistemic theories of justification can be described as trying to identify what features or properties of a belief indicate that the belief in question should be accepted by the epistemic agent (and subsequently used in the steering of behavior, the generation and modification of other beliefs, and so forth). One of the approaches frequently taken to identifying these features is to consider paradigm examples of justified beliefs and unjustified beliefs and try to ascertain what the members of each set has in common. It seems, however, that we cannot simply consider beliefs in total isolation of what else is happening in the agent's cognition. After all, a belief that there is a cup of coffee in front of me at some specified time may very well be justified for me, but unjustified for you.[19] Many traditional epistemic theories, like foundationalism and coherentism, try to assess the justification of a belief by examining its relationships to some (or all, in the case of a holistic approach) of the other beliefs that the agent in question possesses at the time in question. If it stands in the right relationships to the salient set of beliefs, then it is justified, and otherwise it is not.

Alvin Goldman, however, thinks that one problem with many of these popular theories of justification is that they identify a belief as justified without considering what *caused* the belief to be adopted, or what causes it to be maintained.[20] Agents can come to hold beliefs that are true, and that *could have been* justified if they were reached in the right way, but which actually resulted from inaccurate and faulty cognitive processes (like wishful thinking, guessing, or those that are merely emotionally driven). These beliefs, even if they happen to be true, ought not to be

[19] Unless you know me well, in which case you are probably justified in believing there is a cup of coffee in front of me at time *t*, for all values of *t*.

[20] Goldman, 1979/1994, p. 113-115.

seen as justified because their causal history is just not of the right sort.[21] For example, imagine someone who has an adequate grasp of mathematics and basic probability theory such that he *could* realize that his chances of winning the state lottery are extremely slim, but instead comes to this belief purely as a result of being depressed and feeling that "nothing ever goes his way." Even though this individual accepts a belief that is true, and that *could* stand in the right kind of relationships to available evidence or information (and so the belief is *justifiable*), the *way* that the belief was actually formed in this case is problematic, and means that, intuitively, the belief is not justified. A justified belief, on the other hand, seems to be one which does have "good causal ancestry", and results in the right way from the right kind of belief-forming (and sustaining) processes and procedures.[22]

The next step then is to identify what it is for a belief to be formed (or maintained) in the "right way". On the standard assumption that what we are after epistemically is *true* beliefs, it would make sense, so the argument goes, that when we ask whether a belief is justified, what we are really asking is whether the current belief we are considering was formed by a process with a good track record. After all, if the process that formed it is *usually right*, then it stands to reason that it is probably right this time, and a correctly functioning cognitive system should go ahead and accept the belief as true (or at least probably true). This thinking is similar to what is at work when we rely on experts for their opinions or testimony. If we have a botany expert who has so

[21] Indeed, it seems that most, if not all, of the classic informal logical fallacies (*argumentum ad hominem, ad baculum, etc.)* can easily be viewed in this process-centric manner, and this relationship is likely not a coincidence.

[22] This idea also heavily guided Alvin Goldman's development of his well-known causal-theory of knowledge (Goldman, 1967) and his work on "relevant alternatives" (Goldman, 1976 and 1986), which both helped set the stage for his process reliabilist account of justification. These accounts are not discussed here because they were never intended specifically as theories of *justification*.

far been correct in 99% of her identifications of a certain plant, and she then tells us that a particular sample is indeed of that plant type, then, because of her demonstrated accuracy and reliability, we can comfortably assume that the sample is indeed of the type in question.[23] Similarly, if we are equipped with a visual system (thanks to evolutionary selection pressures) which has so far been very accurate in its assessment of whether various patches of yellow in front of us are saber-tooth tigers, we would do well to plan our actions accordingly if our visual processes form the belief in our brains that there is a hungry-looking large tiger two feet away right now. On the other hand, we have probably all wanted to believe that we will win the lottery, will not gain weight from eating large quantities of sweets, and (perhaps) that we could travel through time. However, many (well, sadly, *most*) of the beliefs suggested by our wishful thinking turn out to not be true, and so if we accept the next belief that this wishful-thinking process suggests to us, it seems that something has gone epistemically awry. This is still the case even if the particular belief in question *happens* actually to be true.[24] Even if a cognitive process results in a true belief *this time*, if the process is not typically accurate, then it does not result in beliefs that we ought to believe, and so it seems that all of its resulting beliefs are unjustified as a result of the deficient manner in which they were formed.

So, for Goldman and other reliabilists, whether a belief is justified or not depends on how *reliable* the process used is, and it is obviously this focus on reliability that gives the view its name. Here, then, is an initial formulation of the general idea at work:

[23] It is worth noting that a person can have the necessary expertise to *be* an expert without having her track record assessed and her being *known* to be an expert. In other words, a person (or process) can have a good track record without the track record ever having been checked or assessed in any way. This is an important feature upon which externalist theories build.

[24] If used a sufficiently large number of times, even the most inaccurate process is likely occasionally to generate a true belief.

Provisional Process Reliabilism (1): *S*'s belief that *P* at time *T* is justified if and only if *S*'s belief that *P* resulted from a reliable belief-forming or belief-maintaining process.

This provisional theory requires some important modifications before it can become a serious contender, but first it is worth examining the central concepts in this approach more closely.

Certainly, the everyday notion of *reliability* is no stranger to us. A car is reliable when it starts every day (with very few exceptions), an employee is reliable when they consistently arrive on time for their shift, and a digital thermometer is reliable when it consistently displays the correct temperature (within a specified range of temperatures). It seems that what is common to these cases is that we call something reliable when it has a desirable ratio of good outcomes to bad outcomes, according to some desirable property. Which property is the desirable one, however, depends on the context in which we are discussing reliability. So, let us restrict our attention here to cognitive processes. A few examples of reliable cognitive processes are sense perception (under normal conditions), careful deductive reasoning, memory processes which recall our names when we are asked, and the introspective processes that detect the throbbing pain from a sprained ankle. Examples of unreliable cognitive processes include guessing, wishful thinking, automatically believing something because it has been written somewhere on the internet, or a process that evaluates beliefs solely on the basis of how they "sound" (this last example is from Kornblith, 1980). There definitely seems to be something common to the desirable processes, the reliabilist argues, and this points to which property is desirable. Specifically, what they have in common is that they often produce *true* beliefs and are rarely, if ever, mistaken. In the case of process reliabilism (and most of epistemology), there is a specific assumption made that the good-making feature of a belief is its *being true*. With this assumption

made, and now specified as the context for the reliability assessment, we arrive at the following definition:

Reliability: the reliability of a process is its *tendency* to produce true beliefs. A process is reliable when it produces a high ratio of true beliefs to false beliefs, and unreliable otherwise.

Some questions which immediately present themselves are: what constitutes a "high ratio" of true to false beliefs? Can a process that occasionally makes mistakes still be considered reliable? If so, how many mistakes are acceptable before the process is no longer considered reliable and no longer confers justification on the beliefs it produces? It seems that specifying a non-arbitrary cutoff point will be very difficult, if not impossible. Fortunately, it does not seem that such a rigid division is required in our understanding of reliability because the justification of beliefs does not actually seem to be a Boolean concept that either applies completely or is totally absent. Instead, justification seems to come in degrees, with each belief located somewhere along the continuum from completely unjustified up to completely justified. Only an unusual notion of justification would disagree that a belief that your car will start tomorrow is more justified than a belief that you will win the next two lotteries in a row, but less justified than the belief that adding two apples to two oranges results in four pieces of fruit. If beliefs come in varying degrees of justification, it makes sense that the processes that produce them also have varying degrees of reliability. While there is not a fixed cutoff for counting a process as reliable or not, it is still fairly easy to compare the reliability of different processes to each other simply by comparing the ratio of true to false beliefs that they have produced.

When comparing this ratio of true to false beliefs, it is the *objective* ratio that is relevant to the justificatory status of a belief, which is important since the *perceived* reliability of a

process may vary drastically from the actual track record of the process in question. Indeed, the objective reliability level of a process will in most (or even all) cases not be available to the epistemic agent to whom the process belongs.[25] Certainly, the amount of information that the agent in question has will vary depending upon which cognitive process we are considering, but there will be a great number of processes for which the agent does not have an accurate sense of how reliable the process is (perhaps mistakenly believing that they never fail to identify sarcasm in a friend's tone, for example). Further, there will be many processes of which the agent is completely unaware. This means that an agent might have a belief that is *justified* even if they are unable to justify it verbally or consciously.[26] Since process reliabilism holds that the justificatory status of a belief is not just the result of its relationships to the other beliefs (or other internal relata) that the agent has, it is considered an externalist theory of justification, as it uses states, properties and entities external to the agent as constituents of the justificatory relationship (or as factors in justification—often simply called "J-factors").

Now that we have seen how the process reliabilist understands the eponymous notion of reliability, we will also, of course, want to know what exactly constitutes a "process", given the key role it is playing in the theory. Goldman specifies the term as follows:

---

[25] It is for this reason that Steup (2004) and others suggest that we should instead focus on our *evidence* that a given process is reliable. This, however, results in a non-externalist theory of justification, which may or may not be advantageous. We will discuss this issue further in later chapters.

[26] Because of this, process reliabilists reject the traditional notion of the "JJ Principle", which states that a justified belief in *P* must be accompanied by a justified meta-belief that belief in *P* is justified. Process reliabilism instead holds that having a justified belief in *P* does not entail that one is justified in believing *that* the belief in *P* is justified. It is a consequence of the externalist nature of the theory that an agent may very well not be able to provide reasons or evidence for their beliefs, and thus be unable to access what it is that justifies their belief, but this lack of access does not affect whether, in fact, the belief is justified or not. See Section 2.6 for further discussion.

> [l]et us mean by a 'process' a *functional operation* or procedure, i.e., something that generates a *mapping* from certain states—'inputs'—into other states—'outputs.' The outputs of the present case are states of believing this or that proposition at a given moment. (Goldman, 1979/1994, p. 116, original emphasis)

The definition of 'process' is very broad, and includes things like mathematical functions and computer sub-routines, but the processes that are relevant in process reliabilism are clearly cognitive processes, such as those that take place in a human (or animal) subject's brain, mapping things like sensory percepts, other beliefs, or goals into other cognitive states like beliefs.[27] Given the considerable difficulty neuroscience has had in delineating all the relevant functional neurological processes, philosophers (and many cognitive scientists) tend to formulate their theories at this more abstract level of "processes" or "mappings", so Goldman is not doing anything idiosyncratic here. Ideally, each cognitive process can be described both at an abstract, functional level and at the physical, neurological level, but this is still very much an open question.[28] And, as we shall see in Chapter 2, the delineation of processes poses a serious challenge to process reliabilist efforts.

Before we examine arguments against the classical process reliabilist theory, however, there are some modifications that must be made to give it a fair hearing. The first of these adjustments is made necessary by the fact that many of the beliefs epistemology is concerned with do not result from direct perceptual processes like seeing or hearing, but result instead from what could be called "mediating processes". If an accommodation is not made for these mediating processes, the theory will yield wildly inaccurate assessments of reliability in these

[27] I do not intend to overly restrict 'cognitive processes' only to human or animal subjects; non-terrestrial life or perhaps even artificial intelligence might, in principle, satisfy the relevant conditions just as adequately.

[28] A well-known and influential distinction along these lines is the idea, developed by David Marr in 1982, that there are three levels of cognitive processing analysis, moving from the more general to the more specific: computational, algorithmic, and implementation (Bermúdez, 2010, p.47).

cases. For example, it clearly must be allowed that if someone justifiably believes *p* and also justifiably believes *q*, she is justified in believing *p & q*. As process reliabilism is currently defined, however, we would be forced to evaluate this conjunctive cognitive process purely on the basis of its ratio of true beliefs to false beliefs. This will yield a problematic result if we consider what happens if a person ends up conjoining beliefs where some or all of the antecedent beliefs are false. Since a conjunction is false if one or more of its conjuncts is false, then it could appear that the cognitive process responsible for conjunction is itself unreliable, even though the conjunction operation is perfectly reliable and is not the problematic component that results in the low ratio of true to false beliefs. Indeed, every logic student learns early on that, when starting with false premises, a deductive inference can be carefully and correctly done (thus yielding a valid argument) but still result in a false conclusion.

A related problem can occur when considering the reliability of a memory process. It seems that what makes a memory retrieval process reliable is whether the cognitive process in question accurately retrieves the belief in question. But if the belief that is retrieved happens to have been a false belief, this will, under the current formulation of process reliabilism, negatively impact the reliability assessment of the memory process, since it will now be outputting a false belief. The fact that memory processes operate directly on previously-formed beliefs, which may or may not be have been justified when they were created (and sustained), means that the reliability of a memory process cannot be a function solely of how many true beliefs that process outputs.

To avoid these problematic reliability evaluations, Goldman divides processes into belief-independent cognitive processes (which do not take other beliefs as arguments or inputs) and belief-dependent cognitive processes (which have other beliefs constituting at least some of their

inputs).[29] Belief-independent cognitive processes (like visual perception) are reliable if they result in a high ratio of true beliefs to false beliefs, but since belief-dependent processes cannot have their reliability measured only by their ability to produce true beliefs, their accuracy is instead measured by their "conditional reliability", where, *if all the input beliefs are themselves true*, we then compare the ratio of true output beliefs to false output beliefs.[30] This then gives us the components required for a second provisional formulation of the reliabilist theory:

Provisional Process Reliabilism (2): *S*'s belief that *P* at time *T* is justified if and only if:

1. *S*'s belief that *P* resulted from a reliable belief-independent process, or
2. *S*'s belief that *P* resulted from a conditionally reliable belief-dependent process, and all of the input beliefs into the conditionally reliable process used are justified.

A further modification to this theory can be seen to be necessary by considering cases that involve what epistemologists term "defeaters". In one classic example, if one is looking at a room that one knows to be illuminated by a red light, a plain white sheet of paper will appear to be red. Even though we would normally be justified in believing that a sheet of paper that appears red is a red, non-white, sheet of paper, the knowledge that the room is bathed in red light intuitively defeats the justification that the belief would have had otherwise. Internalists frequently describe a defeater as a piece of evidence or knowledge that provides a "(good) reason

[29] Goldman, 1979/1994, p. 119.

[30] It is unclear whether Goldman here intends conditional reliability to be understood factually or counterfactually. The examples that he provides, and his focus on actual historical track records suggest that we evaluate conditional reliability by examining the subset (which may or may not be a proper subset) of past cases where the input-beliefs have been true. However, other places, especially later work on addressing the "new evil-demon problem" (e.g. Goldman, 1986), suggest that he intends that we evaluate conditional reliability counterfactually, but restricted to the set of "normal worlds". This is discussed further in section 2.1.1.

to doubt" the belief in question.[31] This description, however, includes terms that Goldman think are themselves epistemically evaluative, and so do not satisfy his requirements for an explanatory theory. So what description will a process reliabilist offer? Even though the visual process for color recognition (in a non-color blind subject) is assumed to be reliable, the subject indeed is not justified in forming the belief that the paper is actually red. The reason for this is that there is another reliable process, the one that identifies the colored illumination, which undermines or "defeats" the justificatory status of the belief in question. To get a better understanding of this, let us consider a second example.

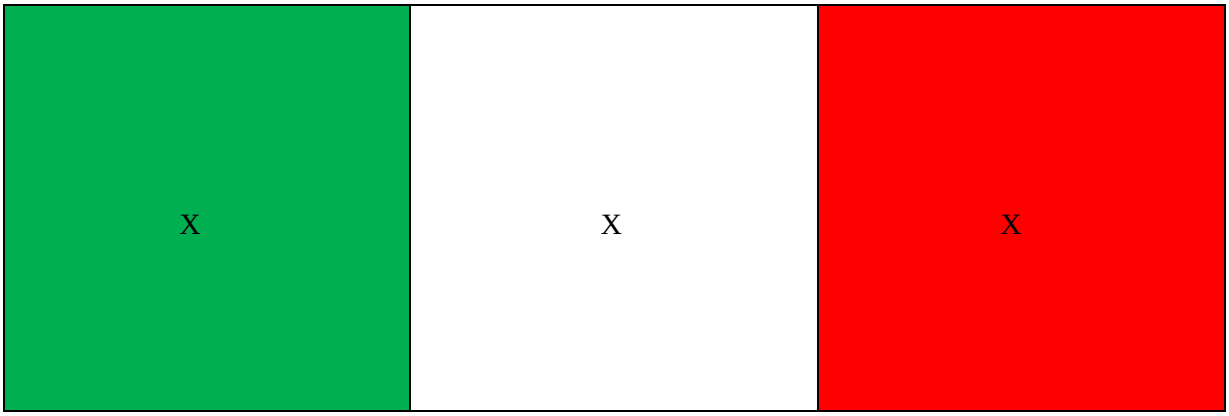

Figure 1 "Colored afterimages due to neuronal fatigue" (adapted from Plate 2, Churchland, 2012).

Another, "easier to see", example is provided by the case of color fatigue.[32] If you focus your gaze on the X in the green square of Figure 1 for 30 seconds and then focus on the x in the middle square, the color of the middle square will appear to be reddish in color for a short time. (If you perform the same experiment but start instead focused on the red square, then the middle

[31] Pollock and Cruz (1999, especially p. 42-45 and p. 195-211) provides extensive discussion of defeaters and the role they should play in an internalist approach. Also see Steup (2016) for additional discussion about defeaters and to what extent they "destroy" the justification otherwise conferred.

[32] Adapted from Plate 2, Churchland, 2012.

square will appear greenish.) At the time that you perceive the middle square not as white, but as reddish in hue, it would seem that you are justified in forming the belief that the middle square is a shade of red, since, after all, normal human color vision is a *reliable* process.[33] So, is this a case where it is just an unfortunate fact that you are justified in believing something false? Well, no, because presumably you have other reliable processes operating at the same time that serve to revoke the justification that would otherwise be conferred by your visual system. For example, you might have glanced at the white box in the middle before conducting the experiment, and so your short-term memory processes might well be retrieving the contradictory information that the middle box is *not* reddish. Or, perhaps you have read about how the neurons responsible for transmitting color information can become fatigued after a prolonged exposure to the same color, resulting in an *illusion* of the color at the opposite end of the neural-encoding spectrum. Once the brain's memory processes retrieve this information, and another, presumably reliable, brain process checks the perception of the reddish square against the memory of conflicting information, the justification for believing that the middle box is a shade of red has been undermined or defeated.

This same general pattern may occur across a wide range of mental processes, involving conflicts between various modalities of perception, memory (both short- and long-term), logical inference processes, and so on. Reliabilism needs to have a component added to the theory that will revoke justificatory status from a belief when there is a different reliable process available that would conflict with the formation of the belief. Once a clause is included that allows for these conflicts, we finally arrive at the full epistemic theory of:

[33] At least, at the level of differentiating green from white from red—even normal, non-color blind visual systems are notoriously unreliable at differentiating very similar shades of the same color.

(Standard) Process Reliabilism: *S*'s belief that *P* at time *T* is justified if and only if:

1. *S*'s belief that *P* resulted from a reliable belief-independent process, or
2. *S*'s belief that *P* resulted from a conditionally reliable belief-dependent process, and all of the input beliefs into the conditionally reliable process used were justified, and
3. *S* does not have a reliable or conditionally reliable process that, if it had been used, would have resulted in *S* not believing *P*.

This is the best-known, standard version of process reliabilism, and will serve as the starting point for the theory to be developed in this dissertation. Goldman has since shifted his attention away from this version of reliabilism, moving instead to work on new theories that are more focused on *social attributions* of knowledge and justification, developing theories that have a lot in common with virtue-based accounts, and developing hybrid theories that combine reliabilism with evidentialism. For example, in Goldman (1993) he proposes that an evaluating agent will consider what psychological processes the epistemic subject in question used to form a particular belief. If all the processes used are on the "list" of *virtuous* psychological processes, then the belief is justified. If any of the processes used are not on the list of psychologically virtuous processes, then the belief is unjustified (or at least "non-justified"). This theory obviously differs somewhat from the standard, classical process reliabilist theory we have been discussing, but even in Goldman's new approaches, what determines whether a process ends up on the list of psychological virtues (or a separate list of psychological vices) is still the *reliability* of the process. So, even with these interesting virtue-themed changes, many philosophers working on naturalist epistemology continue to explore, modify, and hold a version of process reliabilism very close to Goldman's earlier formulation.[34] For this reason, it is this version of reliabilism that will be contrasted with endo-reliabilism

[34] Examples include Grundmann (2009), Lyons (2009), Antony (2004), and Henderson and Horgan (2007).

1.4 The Strengths of a Process Reliabilist Theory of Justification

Process reliabilism has a number of important strengths. In keeping with the naturalists' vision, Goldman aimed to develop a theory that would provide more robust explanatory value than other contemporary theories. Goldman thought that other theories of justification on offer presented analyses of justification that served only to push important questions back a level. A theory which replaces talk of the term 'justification' with talk of being rational or having good reasons, ample evidence, or adequate support (perhaps from a set of "basic" foundational beliefs) may be philosophically interesting, but has not really answered the questions of how and why the belief in question is justified, at least not in a satisfactorily deep way. By using these value-laden terms in the analysis, many questions *about* the evaluative components can simply be asked again using the new terms found within the theory. Goldman's reliabilism, however, attempts to provide more satisfying answers by specifying the conditions under which a belief is justified, speaking purely in non-evaluative terms. A theory that succeeds in offering such an explanation is certainly preferable, *ceteris paribus*.

In attempting to offer this explanation, process reliabilism aims to provide a theory of justification which (1) makes use only of scientific, non-epistemically evaluative terms, (2) is amenable to interrelation with scientific study about our sensory systems and cognitive processes, and yet (3) still retains a degree of normativity by providing an explanation for why someone *should* or *should not* believe a particular proposition.[35] One of the keys to trying to accomplish these tasks is the recognition of the importance of the particular way that a belief was

[35] Concerning (3), in a sense, process reliabilism tells us that an agent *should* hold a belief that results from a reliable process, and *should not* hold one that results from an unreliable process. The kind of normativity offered by process reliabilism is often criticized for being too weak, and is a frequently cited reason for preferring an internalist theory that can better accommodate deontological-style epistemic intuitions. This will be discussed further in later chapters.

formed or caused. Even if an agent has adequate evidence available to support a particular belief, if the agent fails to utilize the information appropriately and instead forms the belief purely on the basis of an irrational whim, then the belief in question is unjustified. Some epistemic theories cannot properly differentiate between beliefs that are justified (in fact) and those that are justifiable (have the potential to be held on justified grounds), and this constitutes a serious problem. For example, a traditional holist coherence theory that holds that the justification of a belief is the result of its coherence with the agent's overall doxastic system cannot draw this distinction (Pollock and Cruz, 1999, p. 79). If a belief is *justifiable* on this theory, this would mean that there are other beliefs available that *could* properly be used to justify the belief in question. But, according to the theory, all of these same beliefs will automatically be included when the belief's coherence is assessed, and so the belief is *justified* if and only if it is *justifiable*. One diagnosis of the problem here is that, without any evaluation of the causal history of the belief, there is no way to separate the beliefs that *could have been* justified (but are not because of the way they were formed, or the evidence they were based on, etc.) from those that are actually justified.[36] Process reliabilism, on the other hand, handles this situation nicely by assigning a central role in the theory to the cognitive processes used to form, modify, or sustain the belief.

The focus on processes and procedures that have a *tendency* to produce beliefs with a certain feature (truth, in the case of Goldman and most process reliabilists) has another advantage. Some philosophers, most notably Descartes, have thought that justification was

[36] This is one advantage of the theory I develop in Chapter 4, since, while it makes heavy use of coherence relations, it also includes the same causal-historical focus as process reliabilism, and so can successfully differentiate between when a belief is justified and when it is justifiable.

infallible— meaning that if a belief was justified then it could not possibly turn out to be false.[37] This justificatory *infallibilism*, however, is problematic because it is unrealistic in its assessment of human cognitive ability. As we all know, humans are capable of making an impressive variety of mistakes. Our senses can deceive us, our memories can fail, we can forget to "carry the one" while doing a math problem, and even our logical reasoning can go horribly awry. If the bar for justificatory status is set so high that it must be impossible for the belief to be false, it is hard to see which, if any, of our beliefs could possibly turn out to meet this requirement. Indeed, even our best supported scientific beliefs fall far short of the infallibilist's justificatory requirements, and so will count as unjustified. Further, the very spirit of infallibilism appears incompatible with scientific practice, since good science assumes that any belief, no matter how well supported by past experiments, evidence, and other justifications could be refuted in the future. For these reasons, most naturalist theories of justification, including process reliabilism, are *fallibilist*, and tolerate the occasional justified belief being false.[38]

Intuitively, most of us tend to think that justification should, in *some* cases, apply to beliefs that are actually false, and having "reliability" defined in a way that requires only that a process *tend* to produce beliefs with the feature in question makes room for some of the beliefs to *lack* that feature. This then allows one to account easily for cases in which a *usually* reliable process occasionally yields a belief that is in fact false, without requiring the process to be denied justification-conferring status—which would seem to be far too strict a standard, since it

[37] A more recent example is Littlejohn (2012), which attempts to argue for an externalist theory where justification is "factive" and no justified belief is ever false.

[38] Most contemporary non-naturalist theories of justification are now also *fallibilist* theories for similar reasons.

would likely exclude most, if not all of our belief-fixing processes, and result in our (nearly) total lack of justified beliefs.

By focusing specifically on *cognitive* processes and procedures and their ratio of true to false resulting beliefs, process reliabilism provides an account of justification in the same kinds of terms as those utilized by sciences like psychology and cognitive science. By dealing with entities investigated by science, the processes and mechanisms that underwrite justification are therefore recast in ways amenable to empirical experimentation. This, obviously, fits quite well with the naturalistic approach to epistemic concepts, and is a large part of the attraction of process reliabilism.

The account of justification proposed later in this dissertation aims to include a similar focus on causal history, process, and procedure, but need not define their reliability in the same manner. Indeed, the theory I propose rests upon a distinction between standard externalist process reliabilism and a version of what could be called "internal reliabilism", which maintains a focus on the reliability of processes, but defines reliability in a different, internalism-friendly manner.[39] First, however, I wish to demonstrate some of the reasons for why I think this philosophical shift is necessary, and so the next chapter will explore some of the problems with the standard process reliabilist theory of justification.

[39] My version of "internal reliabilism" is importantly different from Steup's "Internalist Reliabilism" (2004, 2013, 2016), since Steup keeps roughly the same notion of reliability that classical process reliabilism employs, but holds that it is our internal *evidence of that reliability* (typically furnished by our memory) that actually confers justification on resultant beliefs. As will become clear, I argue that the better route is to *redefine* completely what constitutes the reliability of a process, using relata completely internal to the cognitive system.

Chapter 2: Problems with the Process Reliabilist Approach

While process reliabilism has been quite popular with naturalist epistemologists because of its many strengths, it is not without its share of problems. This chapter presents a range of arguments intended to show that standard externalist process reliabilism is flawed, and demonstrates that its standard formulation should be rejected. If successful, these arguments motivate the adoption of a different theory of the nature of justification, such as the theory of endo-reliabilism I will be advancing in chapter 4. My theory draws upon, and has strong similarities to some aspects of, process reliabilism, and so some of the problems discussed below may also be difficulties for endo-reliabilism. For this reason, they ought to be kept in mind. It should become apparent, however, that endo-reliabilism handles many of these objections more effectively than standard process reliabilism can, and so even if endo-reliabilism does indeed continue to face *some* of these difficulties, it is to be preferred over other available theories.

2.1 Standard Worries for Reliabilism

To begin, it will be worthwhile to consider the most commonly discussed objections to externalist process reliabilism found in the philosophical literature. Though space does not permit me to go into extensive detail here about these objections and the numerous responses that process reliabilists have made attempting to respond to them, a brief discussion of the most important ones is certainly in order. Having the arguments mentioned will be helpful to my

overall argument, as they provide reason to worry about the tenability of the current externalist version of the process reliabilist approach, and will help to highlight the advantages that my theory of justification offers in avoiding them.

2.1.1 The New Evil Demon Problem

Descartes famously raised the possibility that one could have thoughts, evidence, and experiences identical to those that we have, but in fact be the victim of ongoing and systematic deception by a powerful and malicious demon. This would entail that few, if any, of the individual's experiences are veridical, and therefore the resulting beliefs will be largely false. If all of one's experiences concerning external objects like tables, coffee cups, other people, and even one's own hands could be had by someone in a world that contains none of those things, and the experiences would be indistinguishable from the experiences that we have, then it seems, *prima facie,* that we have no reason to be confident that *our* experiences are not themselves the result of an evil demon's trickery. Descartes raised this possibility intending to decisively refute it and thereby defuse what he took to be the strongest possible skeptical objection.[40] Despite Descartes' noble intentions, the evil demon problem has not gone away and continues to rear its head in contemporary philosophical discourse. Indeed, one example of this is to be found in the problem that a modified version of the evil demon problem poses to externalist process reliabilism.

Returning our attention to standard process reliabilism's contention that a belief is justified for an agent if (and only if) the belief resulted from a reliable cognitive process, we might well wonder under what conditions a process must be reliable in order to be justification

[40] A later section of this chapter will discuss in more detail whether such a refutation of skeptical hypotheses like the evil demon problem is available, or even possible in principle.

conferring. It seems that many of our cognitive processes are reliable in *some* settings and applications, but not others, and it has proven surprisingly difficult to ascertain how the line is to be drawn that determines in which environments a process must be reliable in order to confer justification to the beliefs it produces. (Indeed, even identifying how we delineate "types" of environments for analysis is a thorny issue.) One intuitive idea is to attempt to identify justification-conferring, reliable processes as those that tend to yield true beliefs for the subject *in the world the subject inhabits*. This, however, leads us directly into the Cartesian-style objection known as the (new) *evil demon counter-example*, which Goldman describes as follows:

> [i]n a certain possible world, a Cartesian demon gives people deceptive visual experiences, which systematically lead to false beliefs. Are these vision-based beliefs justified? Intuitively, they are. The demon's victims are presented with the same sorts of visual experiences that we are, and they use the same processes to produce corresponding beliefs. (Goldman, 1993, p. 276)

If all of the experiences available to the victims fit the beliefs they adopt, and the victims, let us imagine, are aware of the various epistemic risks and take every available precaution in their reasoning and belief formation, it intuitively seems that their beliefs must be justified. After all, they might be doing exactly the same things, epistemically speaking, that we do on our best days, and we take ourselves to be justified in at least many of our beliefs.

Process reliabilism, however, initially seems to yield an answer that conflicts with these intuitions. If we define reliability relative to the surrounding environment, the evil demon's victim has many woefully *unreliable* cognitive processes, including their perceptual processes, and so any resulting beliefs will be *unjustified*. Various attempts have been made to solve this problem,[41] but here we will only examine two of Goldman's attempts to reformulate process reliabilism to get around this worry.

[41] See Pollock and Cruz, (1999, p. 115) and Goldman and Beddor (2016) for discussion.

Goldman made an early suggestion (1986) that the problem that leads to the conflicting result described above is that, in the evil demon case, we mistakenly evaluate the reliability of the agent's processes relative to the world that the victim *inhabits*. Instead, he suggested that reliable processes are those that are truth conducive in "normal worlds", regardless of whether the agent under consideration inhabits such a "normal" world or not.

> We have a large set of common beliefs about the actual world: general beliefs about the sorts of objects, events, and changes that occur in it. We have beliefs about the kinds of things that, realistically, do and can happen. Our beliefs on this score generate what I shall call the set of normal worlds. These are worlds consistent with our general beliefs about the actual world…. (Goldman, 1986, p. 107)

If we specify a normal world as a world that does include the kinds of external objects that we normally think that we perceive and experience, then any cognitive process that would yield a high ratio of true beliefs *in these* circumstances can be claimed to be reliable. And if the victim of the evil demon's deceptions is arriving at beliefs as a result of these same processes (that are reliable in a normal world), then the resulting beliefs are indeed justified (and thus conform to our intuitions in the evil demon case). However, Goldman encountered considerable difficulty in satisfactorily delineating what constitutes the set of "normal worlds", and many people worry that any attempt to define "normal worlds" is destined to be *ad hoc*, and so unconvincing. Owing to these difficulties, Goldman (1993) adopted a new strategy to solving the problem. He says that his

> basic approach is, roughly, to identify the concept of justified belief with the concept of belief obtained through the exercise of intellectual virtues (excellences). Beliefs acquired (or retained) through a chain of "virtuous" psychological processes qualify as justified; those acquired partly by cognitive "vices" are derogated as unjustified. (Goldman, 1993, p. 274)

In effect, we try to identify which psychological processes are epistemically virtuous, and their excellence is to be found in their truth-conduciveness. Goldman continues, suggesting that

> the epistemic evaluator has a mentally stored set, or list, of cognitive virtues and vices. When asked to evaluate an actual or hypothetical case of belief, the evaluator considers the processes by which the belief was produced, and matches these against his list. If the processes match virtues only, the belief is classified as justified. If the processes are matched partly with vices, the belief is categorized as unjustified. (Goldman, 1993, p. 274-275)

If we presume that our standard visual-perceptual processes *are* included on this "list of virtues", then when we consider a victim of an evil demon, or a brain-in-a-vat, we will find that many of their beliefs about an external world are formed using the same processes that are virtuous, and therefore, justified. Goldman admits that this theory is intended not as a view on what *constitutes* justification, but rather as a theory about when we *attribute* justification to the beliefs of others (Goldman, 2012, p.81). But, as an externalist, Goldman holds that there is one *correct* set of virtues: those that are in fact the most truth-conducive in the actual world, and so he thinks that the virtue and vices approach offers the framework for a solution to the new evil demon problem by identifying which virtues and vices are applicable.

But is this a fair move, to ground the solution to the New Evil Demon problem in "normal" or "actual" worlds? Both of Goldman's avenues of response seem to me to fail to grasp the full force of Descartes' *original* evil demon worry. A skeptic can worry that what we take to be the "normal world" is decidedly not included in the set of normal worlds (as defined), and so to try to specify reliability as reliable in normal worlds is viciously circular since it basically defines reliability as yielding high ratios of true beliefs in worlds where the process yields high ratios of true beliefs. A related objection applies to the virtues-based approach that Goldman has advanced. If the community that is identifying the epistemic virtues and vices is itself the victim of an evil demon's deceptions, then the identification of the virtues and vices is either internalist or is specified in a way similarly circular to the normal worlds approach. Admittedly, the New

Evil Demon problem is a different problem from the Descartes' original version, but I think it is the original is still relevant here for the reason mentioned.

Goldman's approach has another, perhaps more serious, problem. Imagine a possible world, one decidedly abnormal, where the inhabitants have the power to make their wishes come true. In this world, the cognitive process of wishful thinking is highly reliable, since the very act of wishing something to be true *makes* it true. It seems that the fact that wishful thinking is not reliable in "normal worlds like ours" should not mean that the process is automatically deemed unreliable or vicious (and so non-justification-conferring) for those inhabitants in *that* world. So, the challenge of correctly delineating, in a principled, non-*ad hoc* way, in which possible worlds a process must tend to produce true beliefs to count as reliable remains a problematic issue for standard process reliabilism.[42]

2.1.2 Reliable Clairvoyance and Justificatory Defeat

The virtue-driven modification discussed above was also taken by Goldman to address the second major objection commonly discussed in the literature, the problem of "reliable clairvoyance" raised by BonJour (1985), which seems to potentially show that having a belief formed via a reliable process is not sufficient for justification.[43] If we imagine someone, call him Norman, who has incredibly reliable clairvoyance, but has no reason to think that he is

[42] Also see Henderson and Horgan (2007) for their theory of "Tranglobal Reliabilism", which defines the set of relevant worlds as the "experientially possible global environments". Roughly, the idea is that a process is reliable if it has a tendency to produce true beliefs over the set of all possible worlds that are compatible with the kinds of experiences we have. To my mind, the nature of this attempted solution, by suggesting evaluation across a large, perhaps infinite, set of possible worlds, deviates too far from the naturalistic project, and so is not discussed further here.

[43] Lehrer (1990, p. 163) gives a similar example in the case of "Mr. Truetemp", who reliably detects temperatures, and forms true beliefs about them, but arguably still lacks justification for those beliefs.

clairvoyant (nor perhaps even that such a thing exists), it seems that this person can form beliefs via this clairvoyance which should, by process reliabilist standards, be acknowledged as justified (since the clairvoyance is stipulated to be *reliable*).[44] However, as Goldman reports, "BonJour contends that the beliefs are not justified; and apparently most (philosophical) evaluators agree with that judgment" (Goldman, 1993, p. 276). So it seems that process reliabilism will identify Norman's resulting beliefs as justified when intuitively they are not. Using the virtue-driven reformulation, Goldman claims that process reliabilism can yield the same assessment, noting that clairvoyance is *not* on our list of epistemic virtues (and, indeed, makes most people's lists of epistemic *vices*!) Again, however, the virtue-based approach deals with the *attribution*, not the *constitution*, of justification, and so the force of the reliable clairvoyance objection is still being evaluated.

At its heart, the issue here is that process reliabilism has a difficult time accounting for the defeasibility of justification. One might worry that the account that must be offered for why beliefs are defeasible (which must be accounted for in any epistemic theory) ends up being overly complex and seemingly *ad hoc* when addressed from within the process reliabilist framework. If justification is thought to be conferred when a belief *results from* a reliable process, then why is it that the conflict between that belief and another belief can end up undermining one of them? If the central commitment of reliabilism is that processes that are highly accurate tend to yield true beliefs, then it might seem on the surface that an account of defeasibility would also be operating at the level of belief-forming processes. Indeed, as we saw

[44] In fairness, however, Goldman's 1979 account already included a clause for "defeaters" that avoids some worries along these lines. Goldman held that a belief cannot be fully justified, no matter how reliable its forming mechanism was, if one has another reliable process, which if employed, would result in the subject not forming the belief (Goldman, 1979/1994, especially p. 124.)

previously, Goldman's earlier work tried to handle the situation in this way, by holding that epistemic defeat occurs when the subject had a more reliable cognitive process available, such that, if it had been used, would have resulted in the subject no longer holding the belief in question (Goldman, 1979, p. 124). However, in addition to other difficulties with this approach, this suggestion arguably fails to account accurately for the cases of justificatory defeat in which we are most interested. The conflict that is at the heart of defeasibility generally occurs between token beliefs, and not between their processes.[45] Process reliabilists can, and certainly have, offered theories that do acknowledge the role that defeaters play in our cognitive economy (see especially the recent "hybrid" theories developed by Comesaña, 2010 and Goldman, 2011), but these approaches tend to end up forced to rely on something similar to coherence metrics and relations, which I will argue, can be more simply, directly, and easily integrated into other theories (including the one I offer in later chapters).

2.1.3 The Generality Problem

Another major objection to process reliabilism, known as the *generality problem*,[46] asks how the "cognitive processes" that are key components in a process reliabilism account of justification are to be differentiated. It seems that, while any particular belief results from one specific causal process occurring (at a particular time), when we turn to assess the reliability of the belief-forming process *across* different instances, there are numerous ways to group the

[45] There are, of course, some interesting questions that arise about conflicting processes as well, but those issues are largely separate from the issues of one *belief* operating as a defeater for another *belief* that is traditionally seen as crucial to an epistemic theory.

[46] This was also raised by Goldman (1979).

instances and process. This means that the process that formed a belief can be described in any number of ways, resulting in conflicting evaluations of the process' reliability, and thus the belief's justificatory status. "For example, the process token leading to my current belief that it is sunny today is an instance of all the following types: the perceptual process, the visual process, processes that occur on Wednesday, processes that lead to true beliefs, etc. Note that these process types are not equally reliable" (Feldman 1985, p. 159-160).

Indeed, a belief might even count as justified when considered to result from one process type, but unjustified from another. Imagine that I am out walking at dusk, and see what looks like a human in the distance. I then go on to form the belief that there is a person in that location. Unbeknownst to me, it is actually just a small tree with a roughly bipedal shape. Whether my belief is justified will depend on which process is responsible for producing the belief. If we consider the relevant process to be the general process of visual perception, then it seems that my belief is justified because, overall, my visual perception has a fairly good track record.[47] The more specific process responsible for identifying humans on the basis of visual shape has an outstanding track record, thanks to the great many humans I have encountered in brightly lit cities and correctly identified by shape. However, if the correct process to evaluate is my process for identifying mid-sized objects in dim light at a distance, the reliability level falls drastically, and the justificatory status of the resultant belief drops accordingly. And the worst outcome results from evaluating the most specific process: the process of identifying *this* very shape on *this* very evening, at *this* exact time. This particular process has only ever produced one belief, which is false, and so the process is perfectly unreliable and the belief completely unjustified. As we can "see", each of these different processes will have a different track record of accuracy

[47] Well, since my successful Lasik procedure, anyway.

since each process has produced a different set of resultant beliefs. This means that even a single belief will have multiple, incompatible assessments of its justification. This is obviously a serious problem for a theory that intends to present a clear, simple, principled and explanatorily powerful understanding of justification.

It is not clear that there is any way to group different process types in a principled manner. A process reliabilist's tenets may allow the response that the project is simply to employ scientific investigation to discover the natural kinds that already exist in nature at this level, and let this investigation draw the taxonomy for us. This is just conjecture, however, and so the objection from generality is still seen by many as being the most problematic for process-driven accounts of justification. While literature discussing this problem is voluminous, so far no definitive solution has been offered.[48] Indeed, some critics of process reliabilism see this objection as a fatal blow to the theory. For example, Pollock and Cruz label the multiple attempts made by process reliabilists to solve this problem as various forms of "gerrymandering", and suggest that any so-called solution to this problem that is discovered will be unacceptably *ad hoc*, and that it therefore "follows that process reliabilism is essentially bankrupt" (Pollock and Cruz, 1999, p. 116-118).

2.1.4 Swamping/Value Problem

Another problem for the reliabilist approach deals specifically with reliabilist theories of knowledge, but it is the nature of the reliabilist's theory of justification that gives rise to the difficulty. Known as the Value or Swamping problem, this worry centers around the idea that

[48] The extensive discussion of the Generality Problem has resulted in it spreading to theories besides process reliabilism. Bishop (2010), aptly titled "Why the Generality Problem is Everybody's Problem", argues that any adequate theory of justification must contend with the Generality Problem.

holding a true belief that was reached in a reliable manner seems no more valuable than holding the same merely true belief that was not reached via a reliable process. Zagzebski (2003) claims that this is analogous to having a good cup of espresso, one that in a fortunate but unusual series of events was produced by an espresso machine that usually makes dreadful espresso, and then asking whether the delicious espresso would be made any better (or be more valuable) if it had been instead *produced* by a reliable espresso machine that tends to produce good espresso. It seems that it is the goodness of the coffee in the cup that determines its value, and not the manner in which it was produced. Similarly, if an agent holds a belief that is true, it seems that it is the truth of the belief that makes it of value to the agent, and so whether the belief was produced via a reliable process cannot add any value to the belief.

This problem has received a great deal of attention in the literature in recent years, and seems to indicate a problem for the reliabilist approach since it is generally assumed that knowledge is certainly more valuable than merely true belief. But if knowledge is only true belief that is also justified (and presumably meets a suitable anti-Gettier condition), and the value that justification could have added to the belief is indeed "swamped" by the truth of the belief, then the reliabilist theory conflicts with standard intuitions about the value of knowledge.

While my project is not directly concerned with knowledge or its analysis, I tend to think that the value problem does highlight a serious, and in principle irresolvable, problem for reliabilist theories as they are typically formulated. Reliabilists hold that justification consists in being produced by processes that have a tendency to produce true beliefs, and this tendency is grounded in the external fact of whether the resulting beliefs are true or not. This is in place of examining whether the agent has internal *evidence* that the beliefs are true (or something similar), which could involve components that overlap considerably less with the truth condition

for knowledge. As a result, the reliabilist notion of justification starts to seem redundant once we start discussing knowledge, where it is already stipulated that the beliefs under consideration are true. The widespread intuition that *justified* true belief is considerably more valuable than mere true belief seems to me to indicate a need for an account of justification that is not itself directly grounded in truth. Kvanvig (2003) discusses the possibility that the value problem may indeed be applicable to other non-reliabilist theories of justification. As will become clear in later sections, I think that this is correct, and the problem results from the fact that almost all theories of justification assume that justification is valuable *because* it is truth-conducive. This nearly unchallenged assumption in epistemology then produces theory after theory of justification that will be vulnerable to the swamping problem. This is discussed further in Chapter 3, where I argue for an alternative, pragmatic approach.

2.2 Skepticism Undefeated

The new evil demon counter-example mentioned in section 2.1.1 is generally offered as an "intuition pump" aimed at showing that externalist process reliabilism does not fully capture what justification is. When we think about someone deceived by an evil demon, or existing as a bodiless brain in a vat (BBIV),[49] upon consulting our intuitions about whether the victim's beliefs are justified or not, many people find that they intuitively think that the unfortunate individual is still justified in his beliefs, as long as he has made a "best effort" to reason carefully, be attentive to defeaters, etc.

[49] I use 'BBIV' to restrict the discussion away from the "embodied" Brain-in-a-Vat type scenario put forth by Daniel Dennett in "Where am I?" (Hofstadter and Dennett, 1982), which, while interesting, is not really an instance of external-world deception.

Of course, contemporary epistemologists are generally of the view that external world skepticism itself is an ultimately untenable position.[50] While many philosophers will admit that external world skepticism could *in theory* be correct, and maybe cannot be positively defeated, practically it is taken to be unviable. Many people hold that there is simply no compelling reason to accept skepticism; rather, arguments are generally offered to the effect that if a non-skeptical view of the world can be plausibly maintained, the burden rests upon the skeptic to provide sufficient reason to abandon that view.

A variety of other approaches to refuting the skeptic's challenge have appeared in the annals of philosophy. Many of the recent responses to skepticism focus on ways that we could still have knowledge even if skeptical hypotheses cannot be known to not obtain—this includes the well-known approaches by Dretske (1971) and Nozick (1981) focusing on relevant alternatives and the sensitivity condition, as well as work by Williamson (2000), Sosa (1999 and 2000) and Pritchard (2005) discussing the safety condition. This work, however, does not typically engage the likelihood that the external world actually exists.

However, I think that it is important to our theorizing about justification to assess the possibility that we are actually in one of these "skeptical hypothesis" situations. While there have indeed been some arguments provided that attempt to refute the likelihood of the skeptical hypotheses' actually obtaining, I argue in the sections below that these accounts are weaker than they at first seem because they do not fully consider the epistemological implications of highly developed technologies. Some developments in the areas of artificial intelligence, virtual reality,

[50] See, for example, Huemer, 2001; Pollock and Cruz, 1999; and Williamson, 2000. Also, in a paper examining the results from the recent PhilPapers online survey of philosophers, Bourget and Chalmers report that 81.6% of respondents accept or lean toward external-world realism, while only 4.8% accept or lean toward external-world skepticism

neuroscience, etc. seem to have implications that raise the skeptical argument anew, with sharper teeth and a more powerful bite.

While there are more anti-skeptical arguments available than can possibly be addressed here, this section will examine three important classes of attempted refutations. The three particular lines of argumentation considered will be what I label “arguments from intent”, “arguments from probability”, and “arguments from plausibility” (such as those provided by Thomas Reid, G.E. Moore, and Michael Huemer). This section will then demonstrate that newer, skeptically hypothesized world-states (SHWorlds?), ones informed by advanced technology, may prove to be more resistant to these types of argument than were traditional skeptical arguments.

2.2.1 Arguments from Intent

The first attempted refutation of skepticism that I shall consider is what I term the “argument from intent”. While I doubt that many people would find this style of argument convincing against any of the skeptical hypotheses, it is interesting to consider further reasons for its failure against the more recent, technologically driven skeptical hypotheses. This attempt at refuting skepticism took the form of examining the presumed intentions of those beings in a position to effect an external world deception. For example, Rene Descartes attempted to reassure himself that God is, by his nature, such a benevolent being that he would never bring about such a deception. Descartes’ faith in God and God’s benevolence is then utilized to build faith in the reality of the external world of our experience.

It would seem that arguments from intent are predicated on the idea that a being which has sufficient intellect to perpetrate the kind of deception being discussed would, by virtue of

that same intellect, necessary arrive at a certain "correct" moral stance, namely that of goodness. This may or may not be supportable, but even if it is, it provides no guarantee against the technologically powered skeptical hypotheses, for a very simple reason. Technology can be used to provide an "artificial power" to a being who does not innately possess that power. For example, on a daily basis, people send email around the globe, which (usually) arrives within seconds. This is clearly an ability that no technologically unaided human possesses. Further, even knowledge of the functioning of the technology is not required for its use. (Consider email again, where the huge majority of its users operate and utilize the technology without any idea of the binary construction details of the TCP/IP data-packets which are transmitted across the 7 layers of the OSI model.) It seems obvious, then, that any ordinary being from a sufficiently developed culture could perpetrate a complete and consistent deception on another being, with malicious intent, well-intentioned curiosity, or even complete ignorance of the effects of her actions.

By virtue of their technologically driven nature, the more modern skeptical hypotheses seem to involve a shift in focus away from supernatural beings (such as a Cartesian demon, Malebranche's occasionalist God, or Berkeley's God, under some interpretations) to beings who are no longer infinitely insightful as well as awesomely powerful in their nature, but that nevertheless could at least conceivably possess the capability to deceive us about the nature of the external world. As the technology that we use daily progresses, we find ourselves becoming more powerful, and so the ability gap between us and beings capable of deceiving us narrows. When, added to that, one reflects on the actions that certain human beings have been willing to perform in the past, the notion of these same individuals' having the technology available to perpetrate such a total deception is unsettling indeed. In any case, power possessed in virtue of

technology carries no attendant argument against malice, in the way that power possessed as part of the traditional triad of omniscience, omnipotence, and omnibenevolence was imagined to do.

2.2.2 Arguments from Probability

Given the apparent impossibility of definitively *proving* the correctness of either position in the external-world debate, the assessment of the existence and nature of the external world of our experience seems to be relegated to a probabilistic discussion. From this perspective, one of the strengths of the skeptical hypotheses is revealed by the fact that, by the very construction of an adequate skeptical hypothesis, any *a posteriori* evidence that could possibly be had will equally reinforce both external world realism and its corresponding denial. As Michael Huemer puts it in his introduction of the BBIV scenario, "Is there any evidence to suggest that you are a BBIV? No, but if you *were* a BBIV there *wouldn't* be any evidence (at least, not that you were aware of)…Thus, the lack of evidence supporting the BBIV scenario is not evidence *against* it either" (Huemer, 2000, p. 17, original emphasis). Huemer later says that "[n]o matter what observations you make, a suitably modified BBIV scenario can account for them. That is why there cannot, in principle, be any evidence against the BBIV scenario" (Huemer, 2000, p. 17).

This seems to have the effect that any probabilistic comparison will have to be decided on the basis of *a priori* probability (which is informed by only deductive reasoning). One might attempt a refutation of external world skepticism in the following way: notice first of all that there are two possibilities; the external world that we seem to experience either is the way it seems or it is not. Any *a priori* argument assessing the probability of whether it is real will thus start with a probability assignment of .5 for each outcome. The next step in the *a priori* examination is to consider and compare to what extent each outcome possesses each of a set of

virtues that seem to have a certain favorable predictive correspondence. While there is no widespread agreement on the exhaustive set of these virtues, examples of the more commonly utilized virtues that one might employ include: simplicity, consistency, accuracy, and completeness.[51]

Bertrand Russell, for instance, admits that it is logically possible that the world of our experience does not exist external to our minds (in particular, he is considering the possibility that everything we experience is just a dream, but the considerations and arguments he uses are equally applicable to other skeptical hypotheses), but he then attempts to dismiss the skeptical worry based on *a priori* virtues. He says,

> There is no logical impossibility in the supposition that the whole of life is a dream, in which we ourselves create all the objects that come before us. But although this is not logically impossible, there is no reason whatever to suppose that it is true; and it is, in fact, a less simple hypothesis, viewed as a means of accounting for the facts of our own life, than the common-sense hypothesis that there really are objects independent of us whose action on us causes our sensations. (Russell, p. 22-23)

Russell here puts forth two arguments that he thinks defeat skeptical worries. First, he asserts that there is no positive reason to believe a skeptical alternative to external-world realism to be true. This absence of reasons seems intended to function in a manner that shifts the burden of proof onto the skeptic, where it is equivalent to saying that in the absence of defeaters for the external-world-realist position, external world realism is to be adopted. (The notion of defeaters plays a substantial role in Michael Huemer's approach to refuting skepticism, and is considered in more depth at a later point.) As noted above, however, there would not be *a posteriori* evidence if the skeptical hypothesis *did* obtain, so its lack is equally supportive of both the external world realist and the skeptic. Clearly then, this lack cannot be invoked to support the realist position over the skeptical one. This

[51] Later chapters will investigate and discuss theoretical virtues in more depth.

leaves only the possibility of providing an *a priori* justification for believing that the external world does not exist, and Russell asserts that there is no positive *a priori* reason for accepting the skeptical hypothesis. This seems to give both hypotheses, for the moment, an equal *a priori* probability assignment.

Russell then proceeds to claim that, on the other hand, there *is* an *a priori* reason to prefer the hypothesis involving existent external objects, namely its higher degree of simplicity. Russell argues for the increased simplicity of this hypothesis by pointing out that the constantly maintained existence over time inherent in a mind-independent external object provides the simplest explanation of the continuity across actions that we observe in our experience. Russell offers the following example:

> The way in which the simplicity comes in from supposing that there really are physical objects is easily seen. If the cat appears at one moment in one part of the room, and at another in another part, it is natural to suppose that it has moved from the one to the other passing over a series of intermediate positions. But if it is merely a set of sense-data, it cannot have ever been in any place where I did not see it; thus we shall have to suppose that it did not exist at all while I was not looking, but suddenly sprang into being in a new place. (Russell, p. 23)

Thus, Russell posits that an actually existing, mind-external cat can better account for a set of observations of the cat in a variety of states than can individual bits of sense-data, by providing a constant and unified source of that sense-data. This single "source" can then consistently provide a wide range of potential individual perceptions, (for example, cat at different times, in different places, from different angles of observation, by different people, etc.) This is taken to provide a much simpler account of the multitude of experiences we have of the cat (not to mention of the notorious set of counter-factual observations that *might* have happened, but *didn't*) than the "cosmic coincidence" involved in supposing that a large number of

unrelated, cross-temporal and cross-spatial, sense-data all happen to portray the same colors, size, furriness, etc. engaged in a wide range of (potentially causally-unconnected) experiences.

This kind of simplicity move[52] appeals to the highly intuitive line of reasoning that any nomological-type relationship [(x)(Fx→ Gx), or "All *F*'s are *G*'s"] that can account for numerous observed instances ("F_1 is G_1, and F_2 is G_2, etc."),[53] In Leibniz's thought, for instance, simplicity is thought of as unity in multiplicity.[54] Thus, the more that various instances can be subsumed under a single principle, the simpler and more rational the account is taken to be. While positing an independently existing entity to account for a variety of individual sense perceptions is clearly not a case of postulating a law, the similarity in function with respect to deriving simplicity is parallel. The importance of such a rule-like basis for association becomes even more obvious when one considers the addition of a temporal index. An *F* that is *G* at time t_1 could potentially cease being *G* at time t_2 and will therefore require either a unifying postulate functioning to maintain that "*G*-ness" across time, or the conjunction of

[52] Simplicity is notoriously difficult to define, calculate and compare. Simplicity might be a matter of the number of entities, the size or complexity of entities, the number of, or complexity of the relationships between entities, etc. In some of these interpretations, the computer-generated world might easily turn out to be the simpler. There is a wealth of literature on the topic of simplicity and its consequences, including the famous passage in Quine's "Two Dogmas of Empiricism" (1951), which points out that one can use a simpler vocabulary, with the implication that one has to use many words to get anything complex said, or one can use a vocabulary that's simpler in the sense that it makes precision and shortened discussion possible, but it requires that one have all the background requisite to be able to use the words correctly. (The same thing is true of logical systems—one can do everything with Boole's assumptions that one can with Aristotle's, and more, because the former doesn't automatically include existential import with respect to universal propositions—so it's simpler in the sense of making fewer assumptions; Aristotle's logical system is simpler in the sense that it accords most naturally with common sense—if one talks about all the apples in the barrel, one does not have to go the extra step to say that there are some).

[53] For a fascinating treatment of universal relationships, see Fred Dretske (1977), Michael Tooley (1977, 1987) and David Armstrong (1978, 1983, 1991, 1993).

[54] Leibniz (1714)

instances will need to add another, potentially infinite, set of propositions for *each* instance of the relationship.

However, on a skeptical hypothesis, a computer program can account for these universal relationships as well as the more standard "second-order (metaphysical)", or "causal-disposition", etc. approaches recently employed by philosophers; a program can easily be written such that the objects it constructs have certain fixed attributes that will occur, necessarily, in every instance (or member) of that class of objects. The result of this is that a few lines of computer code can establish a nomological relationship between attributes without any of the problems that result from grounding a potentially infinite number of instances in a conjunction of specific cases. This can therefore equally account for the continuity and consistency of our experiences, and may be simpler (owing to the fact that a few compact lines of computer code which only occupy a few "real" objects (like computer memory and hard drives) might still be able to generate the much larger set of objects that we experience as the external world).

Having seen some of the problems with attempting to establish external world realism as preferable based on simplicity, let us briefly evaluate other virtues that one might consider. Presumably, all hypotheses that are not internally consistent would have already been ruled out, and so the remaining skeptical hypotheses are as consistent as is the realist perspective. Both hypotheses provide equally accurate predictions for our experiences and observations, and the accuracy of these predictions and observations in terms of their resemblance to the *actual* external world cannot be considered without begging the question. Finally, both hypotheses provide a sufficient explanation of all the empirical observations that we make. I submit that, in fact, the modern technologically based skeptical hypotheses will rank very closely to the realist

hypotheses on any reasonable virtues that one might seek to employ, and so there will not be any drastic difference between their *a priori* probabilities.[55]

Even if the realist could somehow manage to eke out a slight probabilistic advantage against a skeptical hypothesis using these *a priori* "virtues" (which, as shown above, is by no means certain), the sheer number of skeptical hypotheses will pose a serious danger to the overall likelihood of external-world realism, because the realist alternative will have to be capable of defeating the disjunction of *all the skeptical hypotheses*. A skeptic will wish to point out that the large number of conceivable situations within which we would be drastically deceived about the nature of the external world makes it more likely that our world could be one of those possible worlds. In other words, an increase in the number of possible deception scenarios increases the *a priori* probability that this world is an instance of this deception class. More formally, even establishing the following statements (where *R* is the hypothesis that the external world of our experience is real and *SK* is the hypothesis that the world of our experience is actually created by a computer powering the BBIV scenario, etc.) will not be sufficient for refuting skepticism on probabilistic grounds:

$$\text{Prob}(R) \geq \text{Prob}(SK_{BBIV})$$

$$\text{Prob}(R) \geq \text{Prob}(SK_{VirtualReality})$$

$$\text{Prob}(R) \geq \text{Prob}(SK_{Dream})$$

This is not sufficient to justify a belief in *R*, unless the realist can also establish:

$$\text{Prob}(R) \geq \text{Prob}(SK_{BBIV}) + \text{Prob}(SK_{VirtualReality}) + \text{Prob}(SK_{Dream}) + \ldots \text{Prob}(SK_{N})$$

[55] I think it is important in this evaluation for us all to take extra care to exclude "desirability" as a factor. It is very tempting, perhaps at an unconscious level, to generate arguments against skepticism that employ *reductio ad absurdum* or slippery slope style arguments that presume that what the skeptical conclusion asserts *cannot* be the case because it is not a pleasant alternative.

While the consequences of this should be apparent to anyone familiar with probability calculation, it is worth making the point explicit. Even if one can establish (which again, seems likely to be difficult, if not impossible) that the "reality" of the external world is twice as likely as the brain in a vat scenario, and is twice as likely as the evil demon scenario put forth by Descartes, and so forth, "twice as likely" is woefully insufficient when confronted with any disjunction of these possibilities greater than two in number.

For this reason, the *number* of skeptical hypotheses with which we must be concerned becomes critical. One component that each of these hypotheses must possess for my purposes is that it must boast (at least potentially) a sufficient degree of power, richness, thoroughness, stability, complexity, etc. to be capable of generating a world experience which is indistinguishable from the real world. This condition is intended to exclude hypotheses with obviously detectable defects (such as a being a Bodiless-Brain-in-a-Vat hooked up to an antiquated computer which stalls, freezes, crashes, and can at best generate a 16-color visual experience.[56]) For this reason, all of the technology-based hypotheses under consideration will need to be of a type that utilize considerably more advanced technology than anything with which are currently familiar. As will be demonstrated later in this section, a futuristic thought experiment reveals a situation in which we would not at all be justified in dismissing certain skeptical hypotheses, and this suggests that we are similarly unjustified in ruling out these

[56] While I wish to exclude these possibilities from this discussion, they do, however, prompt what may be a set of interesting questions, namely: if our brains had, from birth, only been exposed to some relatively primitive simulation of an external world, would our brains be capable of identifying any deficiencies in the simulation? Or would they simply adjust such that these experiences would seem completely "normal" and become the standard of reality against which all other experiences are compared? If this is possible, might not there be some medical/electro-chemical process that could be devised which, when performed even on a mature brain, would recast all previous experiential memories in such a way as to make them consistent with the shortcomings of the simulation?

possibilities as currently obtaining, because we may be being deceived into thinking that available technology is far less advanced than it actually is. It is inherent in skeptical hypotheses, after all, that any set of circumstances, at any period in time, which possesses sufficient internal consistency to avoid raising suspicion could be portrayed as the "world".

At this point, one might object by asking: are these really unique thought-experiments, or are they multiple instances of the same class of possible worlds where their differences are not relevant to the debate? I will readily grant that not all skeptical hypotheses are sufficiently different from one another to be added to this disjunction. There are obviously many sets of thought-experiments which would only differ in superficial, non-relevant ways (for example, the computer or device which performs the neural stimulation in a BBIV scenario could vary greatly in its software, hardware, and processes without any variation in the functions performed). There are, however, other differences between some of the technologically based skeptical hypotheses that could very well be critically relevant. As was shown in the previous section, the technology employed in the BBIV case undermines at least one traditional argument against a different form of external world skepticism, and demonstrates that different arguments may well be required if one is attempting to refute different skeptical hypotheses. That being said, an exploration of the precise classification principles required for determining which hypotheses count as novel and which merely as variations is beyond the scope of the current project, and is left as a topic for future research.

Given that technological advance has shown itself capable of generating novel types of skeptical hypotheses, and if technology is capable in principle of infinite advancements (I am inclined to think that this premise might be impossible to establish, even if true, so I will only note its possibility), then it may be the case that there are potentially infinitely many skeptical

hypotheses. Given that only one of the possibilities consists in the external world's being composed of mind-independent physical things, and if it turned out that there were indeed infinitely many alternatives, the *a priori* probability of the non-skeptical hypothesis would be infinitesimally close to zero. Even if it turns out that technology cannot advance infinitely, or that for some other reason only a finite number of skeptical hypotheses are possible, there are already more than a sufficient number of skeptical hypothesis currently available to us to note that the required level of *a priori* probability for the external world realist now seems to be sufficiently high as to be out of reach, barring a radical reformulation.

On the other hand, an external world realist might claim that they can make a similar probabilistic move. After all, the realist position is not just a single hypothesis, either. Because science has furnished us with many different theories about reality, the probabilities of *each* of these should be combined to make a probabilistic disjunction on the realist's side of the comparison. So, consider scientific theories like classic Aether theories, or the more modern String theory, M-theory, and theory of Loop Quantum Gravity. Each of these theories (and many others, of course) claim that there is a real, external world and that it has a certain fundamental nature. And each of these theories has an *a priori* probability of being true, and so their probabilities should be added to reach the total *a priori* probability of external world realism being true. More formally:

$$\text{Prob}(R_{total}) = \text{Prob}(R_{Aether}) + \text{Prob}(R_{StringTheory}) + \text{Prob}(R_{M\text{-}Theory}) + \ldots \text{Prob}(R_N)$$

The realist's claim then would be that, once properly assessed, $\text{Prob}(R_{total})$ does indeed far exceed $\text{Prob}(SK_{total})$, and so much the worse for skepticism.

However, despite initial appearances, I believe that this does not actually undermine the skeptical probability argument offered above. Depending on one's views about the nature of

reduction, one might worry that some of these scientific theories are so radical that it is not clear that they actually *support* the realist's position in epistemology. If, for example, String theory turned out to be true and the "things" that really exist are actually vibrating strings of energy, it is unclear whether we should evaluate beliefs like "my hands exist" as true or not. There are some very complex metaphysical issues here about the relationship between the everyday objects we typically think about (like tables, cars, and kittens) and the fundamental reductive base and the relationships between them. Since this topic is beyond what can be addressed in this section, let us for the sake of argument grant that some solution exists that would make the truth of any of these scientific theories imply the truth of external world realism.

Before moving on from this topic, another related issue hinges on the outcome of the abovementioned metaphysical project. Drawing from Putnam (1981), David Chalmers has argued in his paper "The Matrix as Metaphysics" (2003) that we should be tolerant of surprising reductions of reality even to the point of accepting the BIV hypothesis as a species of external world realism.[57] Chalmers' view is that the hypothesis that we are BIVs is not actually a skeptical hypothesis, but instead a *metaphysical* hypothesis. He thinks that if we indeed are brains in vats, this just means that the fundamental nature of reality turns out to be quite different than we had thought, but does not sow the seeds of epistemic destruction that we might have feared. Chalmers claims that reality, and our epistemic relationship to it, would remain largely intact—it would just turn out that the objects in reality are composed of digital bits instead of subatomic particles and the world was created by machines (or perhaps mad scientists) instead of

[57] The version Chalmers is discussing involves multiple still-embodied brains that are sharing and interacting with the same simulated reality. In the current context, these differences from the classic BIV hypotheses do not matter, as nothing in Chalmers' argument depends on whether it is an individual or many individuals experiencing the simulation.

God or some (random) physical process. This is an interesting, and obviously highly controversial, idea. In fact, it will strike almost all philosophers and non-philosophers alike as completely counter-intuitive. While I would be among the first to admit that there is nothing sacred about intuitions and that when they clash with arguments or theories, intuitions should take a secondary role, it may be worth considering why Chalmers' view would strike so many people as misguided. I would speculate that there are two sources of this intuition. The first is that in the BIV case, the simulation is made *in order to deceive* the subject, while in the case of more standard versions of external world realism, there is nothing to suggest any deceptive intention. I'm not sure that the *intentions* that might go into creating a world are relevant for the metaphysics within that world, but it is possible that an argument could be made for the two having a connection. Either way, few people would embrace something intended as a trick as real. Second, I think intuitions against Chalmers' conclusion are likely motivated by a sense, shared by many people, that the world, whatever it is, must be made of "stuff". Maybe very, very, small stuff, but stuff nonetheless. And if we are BIVs in a computer simulation, then the everyday things with which we feel so familiar are not "things" at all—they are nothing but zeros and ones, nothing but information or data. I suspect that this aspect will be a push too far for many people, and they would reject the idea of labeling a simulated world as real. Of course, this may well mean that our folk intuitions here are misguided. After all, the exact same intuition clash should result when considering more respectable scientific theories like String theory, and its consequence that every "thing" is actually unimaginably small vibrating strings of energy. Anyway, Chalmers makes an interesting case for treating the BIV hypothesis as a realist hypothesis, however, at the least, it seems that the skeptic still retains some formidable

hypotheses with which to aggravate epistemologists, since Chalmers' arguments do not extend to other skeptical hypotheses like the dream or evil demon hypotheses.

So, returning to the topic at hand, the question we were considering is whether a disjunction of scientific hypotheses about the nature of reality can grant the probabilistic upper hand to the realist or not. I think a good skeptic should respond that this does not actually help the realist because, whether we assess the probability of realism as the probability of one single proposition (that the external world exists) or as an aggregate of the probabilities of the scientific theories that entail that the external world exists, in the interest of fairness, we need to evaluate each skeptical hypothesis in a similar way. Consider the bodiless brain-in-a-vat skeptical hypothesis and the simulated world that could accompany it. Some versions of the simulation will involve a simulated reality with aether, some with indivisible particles, others with strings of vibrating energy, and so forth. For any given scientific theory that a realist might seek to include to gain a probabilistic edge, there is an equivalent version of it that is part of the BBIV simulation.[58] As a result, the total probability of the BBIV possibility is a disjunction of the probabilities of being a BBIV in a simulated world with aether or being a BBIV in a simulated world that fits String theory, etc. More precisely then:

$$\text{Prob}(SK_{BBIVtotal}) = \text{Prob}(SK_{BBIV\,+\,Aether}) + \text{Prob}(SK_{BBIV\,+\,StringTheory}) \ldots \text{Prob}(\text{SK}_{BBIV\,+\,N})$$

In order to defeat the skeptical disjunction, the realist needs to be able to show that:

$$\text{Prob}(R_{total}) > \text{Prob}(SK_{BBIVtotal}) + \text{Prob}(SK_{VirtualRealitytotal}) + \text{Prob}(SK_{Dreamtotal}) + \ldots \text{Prob}(SK_N)$$

While on the topic of probability, there is another skeptical argument, adapted from Bostrom (2003), which is worth considering here. If the human race manages to survive long

[58] Remember, the challenge of skepticism is that any empirical evidence for realism is equally evidence for skepticism, and scientific data is no different. Any scientific experiment that provides evidence or confirmation for a specific scientific theory provides equal evidence for each skeptical hypothesis that includes a simulation of that scientific theory.

enough, it seems highly probable that advanced simulation technology will be developed. If that technology were to exist, there is no reason to think that only one simulation would be possible. Instead, it is likely that the technology would be used to run a tremendous number of simulations, perhaps even running simulations of cultures so technologically advanced that the beings in those simulations are themselves capable of running simulations. Given that there would only be one non-simulated reality, but a tremendous number of simulated realities, the chances would be far greater that any particular being exists inside of a simulation than not. Unless there is some reason to think that our species (or others like ours) are unlikely to survive long enough to develop advanced simulation technology, or a reason to think that they would be unlikely to use that technology to actually run simulations, then it seems that it is highly probable that any beings that exist, including us, actually exist inside of a simulated reality! Admittedly, Bostrom's argument is often thought to involve simulated realities populated by simulated beings, and so is somewhat different from the standard BBIV case typically discussed in epistemology. However, a similar point can be appreciated upon realizing that, if a culture of sufficient technological ability were to exist, there is no limit on how many brains might be envatted, or how many different simulated realities created by any one scientist, and there is no limit on even the number of scientists that might be conducting the experiments.

Again, my point here is not to argue that external world realism is false, but to argue that the wide range of possible skeptical hypotheses (and their varying fundamental structures), and the possibility of additional skeptical hypotheses of which we are currently unaware, necessitates that external world skepticism be taken seriously, instead of dismissed as something that "we just know is false". I think that work on epistemic justification should acknowledge that skepticism is

still a live possibility and proceed accordingly, but there is still one more popular style of anti-skeptical argument that should be considered.

2.2.3 Arguments from Plausibility

Arguments such as those presented by Thomas Reid, G.E. Moore[59], and Michael Huemer posit that non-skepticism ought to be preferred on the basis of the fact that it "matches" with our "common sense beliefs". These "arguments from plausibility" seem to be predicated on the success of obviousness as a basis for truth. This seems to presuppose a privileged epistemic position for us where there is a direct and reliable connection between obviousness and actual, independent truth. We must, however, make a conceptual distinction between pragmatic beliefs and true ones. It may be useful to an individual (or a species) to mistake some features of the world, and so we may have evolved certain faulty or inaccurate belief forming methods. For example, the human visual system is capable of experiencing over 2 million colors (Huemer, 2001, p. 95), but in our day-to-day lives, most of that information is never actively registered in our consciousness. This is useful to simplify our perceptions into a less accurate representation because otherwise we would most likely be overloaded by the full experience, and unable to interact effectively with the world. So, given the fact that there is not necessarily a strict connection between even wide-spread human experiences and the actual truth of the matter, initial plausibility might not be the "measure of truth" that we might hope it to be.

Even if there were to be a direct connection between a belief's obviousness and its truth, we ought not to feel comfortable in using this against the skeptical hypotheses. Any competent

[59] As regards G.E. Moore's (in)famous proof: 'I can prove now, for instance, that I do not have any human hands. How? By holding up my two hands and saying, as I make a certain gesture with the right hand, "Here is one computer- generated hologram of a hand," and adding, as I make a certain gesture with the left, "and here is another."'

deceiver would also want to supply us with the impression that (mistaken) "common sense" beliefs are reliable. What better way to maintain the deception, after all, than to create a strong confidence that there is no deception about which to worry? These problems in themselves raise serious doubts about the potential success of arguments from plausibility, but let us put them aside for the moment, and examine more closely one of the more interesting specific arguments offering challenge to the skeptic.

Michael Huemer provides what he takes to be "a general response to any kind of skeptical argument." He says

> [r]ecall that we defined skepticism as any theory that challenges a significant class of common sense beliefs. This means that we cannot, rationally, accept such a theory and also accept those common sense beliefs; we have to choose between them. This being the case, we will, rationally, choose whichever we find more initially plausible. (The "initial plausibility" of a belief is the degree to which it seems true, prior to judgment; in other words, how obvious it is.) (Huemer, 2001, p. 33)

Following from this, according to Huemer, is that "a common sense belief could not be refuted by another, non-common-sense belief; the effect of a conflict between two such beliefs would be that the non-common-sense belief would be refuted instead", and so external world skepticism, as a non-common-sense belief, ends up refuted (Huemer, 2001, p. 35).

Huemer later summarizes the overall structure of his argument as follows:

> 1. Given a conflict between two beliefs, it is rational to reject the less initially plausible one, rather than the more plausible one.
> 2. Common sense beliefs have the highest level of initial plausibility.
> 3. Philosophical theories do not.
> 4. Therefore, given a conflict between a philosophical theory and common sense, it is rational to reject the philosophical theory, rather than common sense. (Huemer, 2001, p. 36)

While several parts of this argument fail to convince me, it will do for present purposes to note only that technologically generated skeptical hypotheses may be capable of completely

obviating this argument's application to external world skepticism. In order for this argument to even be relevant to the external world debate, there must be one more premise added, namely:

> 5. The belief that comprises (or the set of beliefs that comprise) external world realism is a common sense belief

Indeed, Huemer suggests that this is his intent when he says, "A corollary of the conclusion is that, given an argument for skepticism, it is more rational to reject one of the premises of that argument than to accept the conclusion, since the conclusion conflicts with common sense" (Huemer, 2001, p. 35). It is this implied premise (5), however, which may turn out to be untrue.

I contend, instead, that external world realism does not meet Huemer's own criteria for a common sense belief, and so cannot be said to be automatically more plausible. Here are the criteria that Huemer puts forth that beliefs must meet in order to fall into this privileged class:

> i. They are accepted by almost everyone (except some philosophers and some madmen) regardless of what culture or *time period* one belongs to.
>
> ii. They tend to be taken for granted in ordinary life . . . .
>
> iii. If a person believes a contrary to one of these propositions, then it is a sign of insanity. (Huemer, 2001, p. 18, emphasis added)

The hypothesis advocated by the external world realist, however, does not necessarily meet the first criterion, because it may not turn out to be held at all time periods (or, in all cultures, for that matter).

Let us conduct a thought experiment and consider a hypothetical situation, say, 500 years in the future, in which the human race has continued on basically the same track that it seems to be on now. Technology has continued its advance, roughly keeping pace with Moore's law, and doubling in speed every 18 months. As a result, day-to-day technology is incredibly powerful, compact, cheap, and is therefore ubiquitous. Neuroscience has advanced at a similar rate, and now understands nearly every nuance of brain functioning perfectly. One of the new uses that

these incredible advances have been put to is to create small BLVR-visors (pronounced "believer-visors", short for "Brain-Leading Virtual Reality visors") which, when placed on the head, manage to perfectly scan and interpret one's brain states, as well as to stimulate the brain in such a way as to give the experience of a set of sights, sounds, smells, etc. that create a virtual reality indistinguishable from the real thing. While these BLVR-visors were originally developed as games (and likely for secret military use before that), parents had quickly discovered that they could be programmed to function exquisitely well as a distraction for upset children, and as a result, it is now standard practice for newborn babies to be fitted with a BLVR-visor, and to continue to wear one for much of their childhood.

When these children become adults, they will have come to accept perfectly real-seeming (that is, indistinguishable from the world without the visor in terms of coherence and completeness) virtual realities as quite normal, and, being surrounded by exactly the technology required to effect such experiences convincingly, will naturally find themselves very hesitant to accept any particular world of experience, no matter how "obviously real," as actually being the one "true reality."[60]

How does this happen? Huemer details how basic, foundational beliefs about the external world could be defeated:

> The direct realist need not- and should not- hold that perceptual beliefs have a kind of justification that is immune from counter-veiling considerations. He should hold that the justification attaching to immediate perceptual beliefs is, while foundational, nevertheless defeasible justification. The idea here is similar to the legal concept of presumption: perceptual beliefs may be presumed true unless and until contrary evidence appears. As long as there are no special grounds for doubting a given perceptual belief, it retains its status as justified, but when other justified, or *prima facie* justified beliefs start disconfirming it,

[60] One might be inclined to object that these individuals ought to take whichever "reality" they consistently "wake up to" as being the ultimate, true reality. However, it seems that once one has the repeated experiences of presumed realities being mistaken, it will be foolish to put much faith in the certain reality of any experiences that have *not yet* been revealed as mistaken.

> the presumption in favor of the perceptual belief can be defeated and the perceptual belief can wind up unjustified. (Huemer, 2002, p. 585-586)

The above thought experiment demonstrates such "special grounds for doubting," and as a result, these future generations of humans do not share the supposedly "common-sense belief" that the world that they experience is *actually* composed of mind-independent, physical things. Their experiences (especially for individuals who have spent more than half of their lives in virtual reality) function as defeaters of an external world realist belief, and serve to make external world skepticism sufficiently plausible to revoke the justification that Huemer claims for external world realism. This shows that the power of the external-world realism perspective is contingent upon one's experiences, and so may not be as widely and universally held as Huemer's argument requires. Therefore, given Huemer's own definition of common-sense beliefs, his plausibility argument cannot be directed at some of the technologically sophisticated external-world skeptical hypotheses.

Arguments from plausibility also seem prone to suffering from circularity because of their reliance on notions similar to "common sense". For example, in an appeal reminiscent to the 'self-evidence' move adopted historically by a number of philosophers, Huemer specifies in his definition of common sense beliefs that, "They are accepted by almost everyone" (Huemer, 2001, p. 18). The appeal to inter-subjectivity in this case, however, is particularly problematic, given that the existence of other minds is one of the very questions at issue, and so would obviously need to be established first. Agreement by potentially non-existent entities cannot be used to establish and confirm their existence in any way that is not viciously circular. However, Huemer does not have any other option, because any attempt to define common sense using solely internal subjective criteria leaves the beliefs too vulnerable to mistakes (e.g. something absurd could seem *very* obvious to one while on LSD).

This is not the only argument of Huemer's that has appeared historically. In several places, Huemer seems to hearken to Hume, suggesting that the fact that all (non-insane) people, even those who claim to be skeptical of the external world, act in ways that suggest that they do presume a world of mind-independent objects (Huemer, 2001, p. 19 and p. 33). This point is misguided, however, for it fails to take into account that most, if not all, possible skeptical scenarios presuppose some *other* system within which we would find ourselves going about our business. These alternate systems, in order to accomplish their mission, would have to be capable of providing sufficient incentives to account for the skeptic's continuing to act similarly to the external world realist in their shared day-to-day lives. For example, even if one were to somehow be made aware of the indisputable fact that they were a BBIV (for example, maybe the evil scientist intentionally reveals this fact as part of her experiment), one would still have the sum of one's experiences "located" in the simulation, and would be left with continuing one's "life" interacting with it just as before (after the shock wore off, of course…). The system would still provide pleasure based on some choices, and pain based on others, and these are the factors that will continue to determine one's "moves" within the simulation. Unless one somehow has the ability to escape from the deception, one is still going to be bound by the rules that govern that (illusory) reality.

2.2.4 A Realist Rebuttal?

Need technology aid the skeptic alone in the debate? Or can it also contribute new arguments for the external world realist? Our increasing knowledge of the world and its interconnectedness consistently imposes higher minimum requirements on any alternate skeptical hypothesis, because the more we know, the more a deceiver must be capable of

generating a correspondingly larger and more detailed illusion. Additionally, the smaller the objects and increments of time that we can observe, the higher the "resolution" required by the scenario, and its ability to maintain this without "glitches". As a result, it seems that the ongoing consistency of our experience at finer and finer levels of examination would support external world realism.[61]

I think, however, that many of these otherwise powerful objections can be disarmed by applying what I will call the *Criterion of Minimally-Required Deception*, which states that any skeptically hypothesized deception scenario need only to be sufficiently powerful to generate the bare minimum of a world of experience required to fool someone. So, while science tells us these incredible stories of sub-atomic particles in motion and huge intergalactic-expanses that constantly change and evolve, all that a deceiver would need to generate in a deception scenario is a consistent set of headlines in newspapers, sections in textbooks, etc. that *talk* about those discoveries. (If the person [or persons] being deceived happened to look into a microscope or a telescope, the *temporary appearance* of the relevant entities could be generated by the deceiver, but the entities themselves need not exist, or be maintained throughout time and across space.) This all could be accomplished by a simple computer program that monitors the person's status and runs the appropriate subroutines when needed.

2.2.5 Skepticism's Take Aways

Traditional forms of external world skepticism have been a thorn in the side of generations of philosophers, and now, new technologically based variations are providing the

[61] I am indebted to Michael Tooley for presenting this interesting argument to me.

promise of continued challenges to the comfort of common sense well into the future. Technological advancement will bring perfect duplication of reality closer, and will probably eventually permit perfect duplication, unless some unforeseen, untranslatable, or irreducible element is encountered. What will become common in the future when people are surrounded by exactly the technologies (in reality) that the skeptical thought experiments posit? How unlikely will the brain in a vat scenario seem if you spend your days in a medical lab filled with brains in vats? The ability to see, touch and interact with the factors that give rise to skeptical worries may well make those fears seem less far-fetched. (Then again, it might not even take that long—it seems possible that sufficient exposure to science fiction movies that depict these technological deceptions and the uncertainties involved could, on their own, cause one to strongly doubt the reality depicted by the senses.) If we can anticipate that in the future we will have reason to be more skeptical, it seems to me that this also gives us reason to be more skeptical now.

One thing seems certain, as evidenced by the failings of arguments from intent, probability, and plausibility: arguments aimed at refuting the traditional forms of external world skepticism need to be either reformulated, supplemented, or replaced if they are to be effective against the new breed of skeptical hypotheses that are possible with the incorporation of technology. Without stronger arguments available, I think we must take external world skepticism seriously.

If any one of the various skeptical hypotheses does obtain, then there may well be no external world as we experience it, and so obviously any theory of justification that *depends* on the external world as a critical relata in its approach is, if not an outright wrong theory, at least one which is inapplicable and of no use to us. While I am certainly not committed to denying the existence of an external world, the lack of a decisive refutation of external world skepticism

seems to give us good reason to prefer a theory of justification that is neutral about the skeptic's claims.

Of course, some people will think that justification is only important to us at all if the skeptic's worries turn out to be groundless, since one might be inclined to think that if a skeptical hypothesis turned out to be true, then "all bets are off", and our worries about justification will not matter at all in such a "topsy turvy" world. But this is only a preliminary emotional response that, I posit, would quickly pass for any rational being put in such a situation. Imagine that you are being sentenced for some alleged crime (whether justly or not), and the punishment is that you will spend the rest of your life confined within a virtual reality that contains rewards and punishments. I think that anyone will find themselves attempting to ascertain the rules of the virtual environment and trying to optimize their "game play" so as to maximize the rewards and minimize the unpleasant punishments. Even knowing that it is not the "true" actual external world, one still would have a decided preference for cognitive processes that are effective in accomplishing one's goals. For example, there will still be an important differentiation between believing something solely because one wishes that it is true and believing it because of supporting sensory stimulation, and so our usual goal of planning behaviors based on beliefs reached in epistemically responsible ways will still apply.

2.3 Truth Deflationism and Epistemic Anti-Realism

In this section I aim to explore some of Quine's views on ontology and then to adapt some of the arguments famously given by philosophers such as Bas van Fraassen and Larry Laudan against realism in the philosophy of science, and show that they apply equally to a Quinean-style naturalist's epistemology. If, like many naturalists, we think that our normal world

theories about objects like tables, cars, and other minds are simply less formalized instances of general scientific practice, then it would make sense that the same arguments would apply to them as to formalized scientific practice, and this would spell trouble for the reliabilist's approach.

Since the naturalistic epistemological project is so heavily based on Quine's work, it may be worthwhile to discuss in greater detail what he thinks we are doing in our theorizing, both scientific and epistemic. Of particular interest in this section are his views on ontology, since they are relevant to epistemic anti-realism. Quine notoriously holds that "to be is to be the value of a bound variable" (Quine, 1980, p. 15). In his view, our theories provide certain existentially quantified sentences that we accept by adopting the theory containing them, and the variables of these sentences range over a certain set of entities, namely, those that make the sentence of the theory true when substituted into the sentence. As Quine puts it, "a theory is committed to those and only those entities to which the bound variables of the theory must be capable of referring in order that the affirmations made in the theory be true" (Quine, 1980, p. 13). For example, if one accepts a theory of particle physics, it may contain a sentence like "there exists an x such that x orbits the nucleus of an atom is such and such a way…", and the entities that can be substituted into the sentence and result in it evaluating as true are now part of the resulting ontology.

The key question, for the issue at hand, is what Quine means by "commitment" to the entities that the bound variables in our theoretical sentences range over. If this is meant in a sense similar to metaphysical or scientific realists who take commitment to an entity to fundamentally be an assertion that there is an objective fact of the matter, and that in fact, the entity exists, then Quine is intending to make an assertion about the external world. However, this does not fit with the rest of Quine's project. While Quine thinks ontology must be read off of theories, it is the

theory itself that is primary. And given that ontologies are relative to theories, they are not making absolute assertions of any kind.

The key to this view is that Quine holds that all of our observations are theory-laden. According to Quine, every observation that we make depends upon some kind of classification or categorization, in order for us to incorporate it into our existing "web of beliefs". This "web" is composed of all of the beliefs and concepts that we have derived from our experiences to date, which are organized in our minds in some particular, practical way. Given this, no observation can ever be pure observation, as it gets its meaning in terms of the experience of everything else in our experience, including cultural norms of language use, the way our senses work, etc. What someone observes will be shaped by their expectations as well as what theories they currently accept, and so different people can make different observations even under identical circumstances. Because our theories inform the observations that we make, and these observations are organized into the theories that then commit us to a particular ontology, it is the theory itself, and how it is chosen that are of paramount importance for examining our ontological commitments.

Theories serve to organize, categorize, and in many cases, allow us to make predictions about the future; but how do we decide which theories to accept? For Quine, this choice is determined by pragmatic considerations. The worth of a theory is chiefly found in how usefully, adequately, and simply it explains past experiences and allows us to make future predictions concerning a matter that is important to us. For example, depending on which of one's purposes and interests are most pressing, one may equally adopt a naïve folk conception of the world or a sophisticated, philosophically informed view, as each world-view is better suited for certain purposes. In fact, Quine even goes so far as to encourage a pluralistic approach to choosing

theories and their resulting ontologies, where each theory, with its constitutive ontology serves a pragmatic purpose, saying that "the obvious counsel is tolerance and an experimental spirit" (Quine, 1980, p. 19).

So, while Quine holds that our theory selection is a purely pragmatic affair, once we have accepted a theory we will, in so far as we are reasonable, *act* as if the entities entailed by the resulting ontology exist, and not merely as if the theory were "make believe". He says that "we can never do better than occupy the standpoint of some theory or other, the best we can muster at the time" (Quine, 1960, p. 22). However, given our epistemic limitations and the ensuing impossibility of ever viewing the world from outside of some theoretical framework or "web of beliefs", we must recognize (if we wish to be epistemically honest, of course) that these existential claims are only *tentative*. As Quine says, "a conflict with experience at the periphery occasions readjustments in the interior of the field [web of beliefs]" and when these readjustments occur, "no statement is immune to revision" (Quine, 1980, p. 42-43). It may turn out at any point that we make a new observation that requires us to revise our best theories, and these revisions could entail completely different ontological commitments.

When one of these revisions does change our ontological commitments, this is not a realization that the entities previously postulated were adopted in error, or that we suddenly discover a phrase like "phlogiston exists" is false. For example, Quine says that while his best theories commit him to an ontology containing physical objects and not Homer's gods, "in point of epistemological footing the physical objects and the gods differ only in degree and not in kind" (Quine, 1980, p. 44). Instead, it is the theory that changes, hopefully becoming better: simpler, more useful for making predictions, better able to explain various phenomena, and so forth, and this change results in changes to the divisions and categories into which we organize

our experiences. In this way, our commitment to entities, while entailed by our theories, is not an external commitment to the correctness of the theory or its ontology. So, according to Quine, even when one accepts a statement concerning the existence of an entity, the statement cannot be accepted as objectively true. When an entity results from our web of beliefs, we do "believe" in that entity's existence, but only within our present interests, purposes, and particular theory aimed at achieving them, and this belief is not accompanied by a corresponding belief about the web of beliefs itself, except in so far as we believe that our current theory is the *best* theory we have for those particular interests. This tentative commitment to entities because of the value of the theory that entails them for one's present purposes paints a very different picture from the more realist assumptions that most externalist process reliabilists endorse. If Quine's picture of our epistemic theories and the resulting entities they describe is correct, then is doesn't even seem like our webs of belief include the kind of claims that a process reliabilist account of justification is trying to evaluate.

Making use of related insights, Bas van Fraassen presented a now (in)famous argument in favor of scientific anti-realism. According to van Fraassen, the best that we can hope for in science is to arrive at a theory that is completely empirically adequate, and there is no legitimate, non-arbitrary way to take the further step to asserting that the theory is *true* because there will be many, if not infinitely many, theories that are equally empirically adequate and we would have no principled way of deciding between them to identify the "true" one. However, scientific practice is obviously alive and well, and so it seems that we ought not to view scientific practice as even aiming at truth.

A similar conclusion might be drawn from one of Larry Laudan's arguments (1981), which Stathis Psillos eloquently summarizes as follows:

> [t]he history of science is full of theories which at different times and for long periods had been empirically successful, and yet were shown to be false in the deep-structure claims they made about the world. It is similarly full of theoretical terms featuring in successful theories which do not refer. Therefore, by a simple (meta-)induction on scientific theories, our current successful theories are likely to be false (or, at any rate, are more likely to be false than true), and many or most of the theoretical terms featuring in them will turn out to be non-referential. Therefore, the empirical success of a theory provides no warrant for the claim that the theory is approximately true. There is no substantive retention at the theoretical, or deep-structural level and no referential stability in theory-change. (Psillos, 1991, p. 101)

This "pessimistic induction" points out that if we look back over the history of scientific practice, we find that most, if not all, previous theories have now been shown to be false (and have thus been relegated to the scientific "graveyard"), and so it is probable, given induction, that all of our current scientific theories are also false. But even if this is likely, our scientific practice still seems to have value, and so arguably the value must be found in something other than whether the theories it develops are true.

If these arguments succeed, and given the similarity of formalized scientific practice to our general informal scientific practice in the day-to-day world, they seem to indicate that our epistemology should instead be aimed at empirically adequate and empirically successful beliefs, rather than at truth, and that some version of epistemic pragmatism is the best that we are likely to be able to achieve. I will explore this aspect further in section 3.3.

I also wish in this section to raise a worry about what reliabilists even mean by "truth". The standard correspondence theory of truth is treated commonly as a relationship between *only* two domains (for example, 'snow is white' is true if and only if actual snow is actually white.) On this view of truth, a true belief is one where the subjective state (believing that *P*) corresponds with the relevant objective states (states of affairs, matters of fact, etc.). Unfortunately, however, as Kant argued for extensively, we are actually forced to contend with *three* domains. We have the world of our internal representations (or our "thoughts"), the *phenomenal* world that our thoughts represent as external to us (the "target" or referent of our

thoughts about external things), and then the *actual* external things, or "noumena", that are the "things in themselves" *without* any subjective interpretation or experiencing.[62]

This distinction is easy to see if we take the skeptic's worries seriously (as I think we must if we wish to operate in good epistemic faith). If one imagines that one is being deceived by an evil demon, is currently dreaming, or is a BBIV, the three domains for possible correspondence relationships become quite clear. For simplicity, let us consider the BBIV case. First, there is the domain of the individual's thoughts: one's beliefs, one's desires, thoughts about one's self, and so forth—all of which are purely internal to the victim's consciousness. Second, there are the phenomenal appearances of an external world that are presented in the virtual simulation, which can certainly differ from the victim's thoughts about them. (For example, the victim might misremember and believe that her "car" is parked at the "north" end of the "parking lot", while it was actually parked at the "south" end.) There is still a meaningful discussion to be had about the correspondence between these first and second levels, between the victim's beliefs and the phenomenal appearances (experienced as external to the victim). The third domain then will be the level of *actual* objective reality, at which the victim has no hands or a body, but

[62] Many externalists may be inclined to reject this categorization, as they would prefer to see the phenomenal world that our internal thoughts represent as *being the same world* as the external noumenal world. However, while we all *feel* like this is the case in our day-to-day experience (and of course it *would* feel that way, in principle), and all want it to be the case, this line of thinking assumes (without decisive argumentative support) that there is no chance of a skeptical hypothesis obtaining, since it assumes that the things we *represent* as external to ourselves are actually the external things themselves. As we should all know from our nightly dreams, however, we can very well experience flying toaster ovens, giant talking squirrels and shining white unicorns as being external to us in our sensory experience, with these entities clearly seeming to be distinct from our internal minds and "out in the world", but as soon as we wake up we come to realize that they did not *actually* exist in the *real* external world. Kant's distinction provides us with language to discuss *whether* the phenomenal world and the noumenal world are the same or not, and points out that if our internal representations are indeed about the actual external world, this needs to be established and not just taken for granted. Simply put, it is conceivable that our experiences could be identical to what they are and yet *not* be about the noumenal world at all.

instead is only a brain in a vat in a laboratory somewhere. At this level, any beliefs the victim has about her car are obviously false since the victim's car does not, and has never, existed.

Often when we claim a thought or sentence is "true", we mean that our internal representations correspond to our actual sensory representations that seem to *represent* the world that is not internal to us, or, to put it differently, the world that is "not me". (Here it seems to me that we mean something very close to an assertion that the internal representation in question is completely empirically adequate— no possible test will yield conflicting sensory results, the representation can be inter-subjectively verified, etc.) But we also sometimes, especially when engaging in epistemic theorizing, treat our supposedly "true" beliefs as being true in virtue of the correspondence between our internal states and the *noumenal* objects that exist independently. The problem is that the noumenal objects and our representations of them are *not* the same thing, and so we have no license beyond wishful thinking to conflate these two potential correspondences. Some people have, of course, attempted to license this inference as an inference to the best explanation, or in some other way, and this is obviously an important and interesting project. The worry I wish to raise is that many epistemic theories, including, I think, externalist process reliabilism, seem plausible only because they are insufficiently rigorous in distinguishing between the two correspondences, and often do not specify which they mean when they talk about "truth".

2.4 Lacking Holism and Running Afoul of the Quine-Duhem Thesis

Let us now turn our attention to another worry that the ontological commitments entailed by externalist process reliabilism do not quite fit with the naturalist charter. Much of the naturalistic epistemological movement was motivated by Quine's three tenets of theory-laden observation (that every observation is influenced by the theories the observer currently accepts),

under-determination of theory by observational data (that every observation has numerous theories with which it is compatible), and the impossibility of isolating any hypothesis from background assumptions for direct testing—commonly known as the Quine-Duhem Thesis.[63] As we have seen, the Quine-Duhem Thesis holds that no scientific hypothesis can be tested in isolation from the rest of the theory of which it is a part, and so can be neither empirically confirmed nor falsified. If one is considering a hypothesis, for example "All copper conducts electricity", the only way to test the hypothesis is in conjunction with a tremendous number of background assumptions such as that one is able to reliably identify copper, that one's instruments used for detecting electricity are functioning correctly, and that copper's conductivity is not influenced by the color of shoes worn by the experimenter on Tuesdays. A negative experimental result may indicate a problem with the hypothesis, or it may indicate a problem with one or more of the background assumptions. A positive result may be evidence for the hypothesis in question, or it may be that a mistaken background assumption is "saving" a bad hypothesis from being disconfirmed. For example, a chemist might have an hypothesis that two chemicals will generate heat and bubbling when mixed in isolation. When the experiment is performed, the predicted reaction does indeed occur, but actually results from one of the chemicals being mixed with left over residue in the beaker from a previous experiment. The false background hypotheses that "the experimental equipment is clean" and that "there are no other chemicals present" could save the false initial hypotheses from falsification. Additionally, every observation is theory laden, because what one observes varies depending on the background assumptions one makes when performing the observation.[64] These tenets seem to some of us to

[63] See Quine 1970, 1980, 1990, Psillos 1999, and Antony 2004 for discussion.

[64] An interesting exploration of this and its ramifications for scientific practice can be found in Okruhlik 1998.

paint a picture where we are forced to restrict the grounding of our "web of beliefs" to internally accessible measures, such as the coherence between our beliefs, constrained by the other theoretical virtues. On this Quinean view, tables and chairs are part of our current world *theory*, and the Quine-Duhem thesis restricts language, beliefs, and evaluations of the beliefs about these everyday "theoretical objects" to being coherent, parsimonious, explanatorily powerful, etc. at best. With this holistic picture in mind, it is hard to see how externalist reliabilism, which ties the reliability of a process to the objective truth of its *individual* resultant beliefs, can be accommodated by naturalist commitments.

2.5 Not Useful for Regulating Belief

While the debate between internalism and externalism is too large and multi-faceted to be examined here,[65] we should consider one more of the classic objections that internalists frequently raise against externalist theories of justification. According to this objection, externalist theories are flawed because they are not useful for an epistemic agent, since they do not allow the right sort of self-regulation of belief adoption.

As we have seen, one of the classic ways of drawing the distinction between internalist and externalist views of justification is by considering someone deceived by Descartes' evil demon. Internalists think that, since the victim's experiences would be indistinguishable from our own and we think that our resulting beliefs are justified, the victim is also justified in adopting the beliefs resulting from their perceptual experiences as long as they have made a best effort at being properly epistemically rigorous. And why might we be inclined toward this internalism? Most internalists believe that a correct theory of justification *must* allow an

[65] See Kornblith, 2001, and BonJour and Sosa, 2003 for discussion.

epistemic agent to consciously apply it in a regulatory capacity in the day-to-day activity of evaluating potential beliefs. This is to say that the correct theory of justification, whatever it turns out to be, must be usable by an agent when trying to decide which beliefs to adopt, which to reject, and possibly how much credence (or degree of belief) should be granted to particular adopted beliefs.

Externalist process reliabilism is ill-suited to the performance of this regulatory role, because it seems that in order to allow an agent to decide between competing beliefs, the agent would need objective access to the independent states of the world in order to first ascertain whether a particular belief-fixing process is, in fact, reliable. This access, however, is just the kind that no human subject can possibly have, because every agent is epistemically limited by their subjective perspective to their introspectively accessible internal mental states, perceptions, experiences, and so forth. This inclines many philosophers to think that if justification is to be regulatory, it must be "built" out of these internal states that subjects *can* access and utilize in their epistemic decision-making.

On a related note, it is thought by many internalists that only an internalist theory can be normative, since only if our theory allows for self-regulation does it give us the opportunity to "choose" to do what we *ought* to do. This is related to one of the central slogans related to normativity (which, like almost everything in philosophy, is somewhat debated), that says that "*ought implies can*". If one *ought* to pursue only holding justified beliefs, it would seem that one *can* pursue this goal and, in theory, attain it, and this seems only to be possible if *all* of the information relevant to justification is internal to the agent and can be evaluated.

2.6 JJ Principle

Depending on one's views, this item may or may not be a problem for the standard version of process reliabilism, but either way it serves to highlight a difference between standard process reliabilism and internalism. The difference is that, while many internalists are inclined to accept it, process reliabilists are forced to reject what is called the *JJ Principle.* Simply put, the JJ Principle, or more descriptively, the "J→ JJ Principle", claims that if someone is justified in believing a proposition, then it must be the case that they are also justified *in believing that* they are justified in believing the proposition in question.[66] Since at least some of the factors that serve to justify someone in believing a given proposition *P* are external to the individual, there could, and almost certainly will, be cases where the individual is justified in believing *P*, but is unjustified in believing *that* they are justified in believing *P*. For example, we can imagine autistic savants like Dustin Hoffman's character in the movie *Rain Man* (1988), who sees matches fall on the floor and almost instantly forms the belief "there are 246 matches on the floor". It may be that this count results from a highly precise and very reliable brain process, and so the individual is justified in the belief, but could very well have an unreliable process that assesses this belief forming processes. Goldman also gives the example of a young child who will have many first-order beliefs formed by reliable processes, but will not yet have developed all of the neurological processes required for the higher-order assessments.

Still, internalists will often think that the JJ Principle is correct, since if one can consciously access the reasons or evidence that adequately support a belief, the very possession of this support itself serves to justify one in believing that one is justified. So, while not everyone

[66] This closely mirrors the (better-known) "KK Principle" much discussed in the literature on knowledge, which asserts that if *S* knows that *P*, then *S* also knows *that S* knows *P*. See Williamson (2000), among others, for discussion.

is committed to the JJ Principle, those that find it intuitively correct will see reliabilism's rejection of it as yet another problem.

It also worth noting that an externalist version of the JJ principle might be possible to construct. One might think that in order for a belief in *P* to be justified, it must result from a reliable process and there must be another reliable process that has generated a belief that the process that produced *P* is reliable, and so on. However, even this externalist-friendly version is likely to be rejected by process reliabilists, for at least two reasons. First, it seems likely that many processes, even if in fact they are highly reliable, will not be accompanied by the requisite higher-order belief, and so too many beliefs that should come out as justified on the theory will fail to do so. Second, a serious problem is likely to result from what appears to be an inevitable infinite regress. In order for a first-order belief to be justified, there must be a second-order belief about the reliability of its formation process, but there must then be a third-order belief about the reliability of the second-order process, and so on.[67]

Although I do find the JJ Principle tempting in some ways, overall I side with process reliabilists in rejecting both versions of it. If we are aiming at a naturalistic theory, I fail to see how we could realistically and fairly place the demands that accompany the JJ Principle on our cognitive systems.[68]

[67] See Goldman (2012, p.73) and Goldman and Beddor (2016) for further discussion.

[68] I think that it is possible that the theory that I develop in Chapter 4 might be capable of being adapted in such a way that it could satisfy a more limited version of the JJ Principle (where, for example, a first-order belief requires an accompanying second-order belief of the right kind, but does not require a third-order belief). Because I locate the relevant justificatory factors as completely internal to the agent's cognitive system, I suspect that I am in a better position to meet even this restricted version than standard process reliabilists are. This would need to be explored further, and given the dubious nature of the JJ Principle, is not of critical importance to the current project.

2.7 Conclusions

While the literature on process reliabilism, its problems, and possible modifications to it is voluminous, the issues discussed above should serve for our present purposes.[69] The objections to standard process reliabilism discussed in this chapter were raised for two reasons. First, having these objections in mind will allow us to highlight some of the advantages that my new theory of justification offers, since I think my theory can solve some of these objections outright, and because of its different structure has a better chance of having solutions to the others worked out at a later point. Second, and more generally, I hope that this chapter has shown that naturalists cannot simply rest on their epistemic laurels—there are quite a few very thorny issues facing process reliabilism, and definitive solutions to them do not appear to be forthcoming. This fact should convince us that the development and exploration of other theories of justification like the one to be developed in chapter 4, is not only worthwhile, but essential to the naturalist agenda. Naturalist philosophy is too important of a project to tie all of its epistemic hopes to only one theory, especially one so riddled with difficulties.

[69] Goldman (2012) and Goldman and Beddor (2016) include further discussion of these and other topics pertaining to process reliabilism, and is recommended to the interested reader.

Chapter 3: Lessons Learned—Constraints on a Better Theory of Justification

The arguments discussed in the previous chapter signal a need for considerable modifications to the standard process reliabilist account of justification. Not all of process reliabilism's approach should be abandoned in the face of these objections, however, for as I argued in Chapter 1, the commitments to the methodology and ontology of naturalism and the focus on the causal historical process by which an agent adopts or maintains a belief are critical to a successful and informative account of justification. The current chapter will identify two additional *desiderata* that the new theory of justification on offer in Chapter 4 aims to satisfy. Specifically, the theory should be *internalist* in nature, although I will argue for a somewhat non-traditional notion of what counts as an internalist theory, and the theory should be properly constrained (or "bounded") by our epistemic situation and be completely pragmatically oriented.

3.1 Internalist Theories of Justification

Given that many of the objections we just considered were targeted squarely at the externalist nature of standard process reliabilism, it seems beneficial to move away from externalism and instead adopt an internalist approach to justification. While I will not here undertake a thorough review or classification of internalist theories, a quick description of some of the main internalist views will help to shed light on the version of internalism that I am advocating. Justificatory internalists hold that the justificatory status of a belief is a function of

nothing but some set of states that are *wholly* internal to the epistemic agent. Contemporary internalists tend to restrict these internal states either to reflectively accessible mental states (*accessibilism*) or else mental states more broadly construed (*mentalism*).[70] In the next section, I will argue that accessibilism is not an option because the "reflective access" requirement excludes many of the states essential to conferring justification. The plausibility of mentalism depends on how it is understood. If self-proclaimed mentalists are willing to include the kinds of low-level states that I will be discussing, then I am happy to call myself a mentalist. But if, as seems to be the norm in the literature, mentalism only slightly enlarges the set of relevant internal states from consciously accessible states by including other high-level mental states like hopes, fears, and desires, while continuing to exclude the lower-level states, then it will suffer the same problems as accessibilism. Still, internalism's increased compatibility with allowing justification to perform a regulatory role, its "common sense" nature and consistency with the intuitions in both the new evil demon problem and the cases of Norman the Reliable Clairvoyant and Mr. Truetemp, as well as its immunity to the Swamping problem all warrant efforts to modify and improve the internalist approach rather than jettison it.

3.2 Shifting to Subdoxastic Internalism

The theory of justification offered later in this dissertation is indeed internalist, because it holds that justification results only from sets of relata internal to the epistemic agent (and that therefore all the J-factors are internal). Standard access internalism, however, will not quite do for my purposes here. The problem that needs to be addressed is that an internalist account of

[70] The paradigmatic contemporary description of accessibilism, or access internalism, is found in Chisholm, 1988, while Conee and Feldman are widely credited with the development of mentalism, especially in their 1985 and 2004.

justification that relies heavily on coherence relationships, as does the theory I will offer in chapter 4, may not have adequate resources to do all of the justificatory work required of it.[71] How can something like a belief that I am seeing a table in a room I have just entered for the first time be justified solely by appealing to its coherence and inter-relation with other *conscious* beliefs that I possess? After all, as has often been objected to coherentist theories, there are likely to be multiple incompatible beliefs that are all *equally* consistent with my other background beliefs.

To address this, I follow John Pollock and Joe Cruz (1999) in defining internalism in such a way that it incorporates internal *subdoxastic* states as internal relata that can participate in justifying our beliefs. For example, on an approach to internalism that makes use of subdoxastic states, if my belief that there is a table in this room is justified, its justification will depend heavily on what is taking place within my visual cortex, my optic nerve, and along my retinal cells, and not just my other *beliefs* about rooms, tables, and so forth. This view squarely rejects what epistemologists have called the "doxastic assumption"— that only beliefs can be relevant to the justificatory status of a belief.

As we proceed, we will see that if we wish to talk in terms of accessibility and yet keep pace with scientific discovery, we should replace the requirement that the determiners of a belief's justificatory status be "accessible by the subject" with the requirement that they be "accessible to the subject's *cognitive systems"*. This will allow things like each neuron's activation state and trajectories to function as components of the justification relationship.

These scientific posits are still importantly different from the states of affairs in the external world that I have argued should not be allowed as J-factors, because, as internal

[71] See BonJour (1985) and Pollock and Cruz (1999) for discussion.

components of the cognitive system, neural activation states and the like *can* be directly accessed, utilized, and (potentially) modified by the system, all within the system's bounds. Perhaps even more importantly, even as a skeptic trying to make minimal theoretical commitments, I must stand on some "planks" to be able to work on the others, just like repairing "Neurath's ship" while at sea (Quine, 1951). Given that the project of epistemology fundamentally and centrally involves epistemic states (such as beliefs) and the systems that produce and maintain them, both of them are indispensible to the project. So the question then is how they should be understood. Our best current scientific theories tell us that the "self" or "agent" that epistemology has always investigated is best treated as a physical system composed of neurons and their interrelations. Of course, the scientific understanding of the self could turn out quite differently, but as things stand currently, and as a naturalist, I think it is legitimate for me to help myself to these states for use in my theory. Admittedly, these states are not traditionally seen as internally accessible, because they exist below the level of conscious introspection. Restricting internal accessibility to *conscious* access in this way, however, seems overly limiting, misguided, and to fail to heed the findings of science.

Internalists, both past and present, tend to stop their analysis of justification at what had previously appeared to be an intuitive point. If we are interested in when an agent has formed beliefs correctly, it might make sense to look for the lowest level of accessible data and describe the relationship between that data and the resulting beliefs. As we saw previously, Bertrand Russell (1912) and others identified this lowest level as "sense data" and derived an epistemic system using these data as the foundational building blocks for the rest of an agent's beliefs. More recently, as we've seen, some authors have focused on "seemings" as the proper grounding for our beliefs. Consider again, Michael Huemer's Phenomenal Conservatism, which holds that

"[i]f it seems to *S* as if *P*, then *S* thereby has at least prima facie justification for believing that *P*" (Huemer, 2001, p. 99). This fits quite well with the internalist project, but stops the search for the full causal and supportive story too soon. The reason seemings appear to be such a good candidate for the "reduction" or analysis of justification is that they are *already* epistemically value-laden. I think the same point can be made about internalist theories such as evidentialism (Conee and Feldman, 1985 and 2004) and internalist reliabilism (Steup, 2004, 2013, and 2016) that instead employ *evidence* and the act of recognizing that something provides evidence.[72] Both seemings and evidence (as typically used at a doxastic, consciously accessible level in these theories) are the results of a long chain of what I take to be the actual epistemically interesting processes.

If we consider what happens when Barack gazes on Michelle's face and forms the belief that "That is Michelle Obama", it seems plausible that Barack comes to hold this belief because it just "seems" to him that it is her face that appears in his visual field. This type of analysis made sense previously, but I maintain that, in light of our recent scientific progress in understanding what is taking place in the brain, identifying a conscious state such as a seeming as the fundamental basis upon which our beliefs are justified is mistaken and *ad hoc*. Our best scientific theory of how Barack identifies Michelle's face does not bottom out in a seeming, but instead also includes a rich and complex story of how Barack's visual system operates, how his facial recognition neural networks have been trained and configured by experience, how these systems are integrated into the other systems in his brain which contribute to his belief

[72] Steup's theory is somewhat less vulnerable to this objection, thanks to the role that the reliability of processes plays in it. This allows the theory some contact with and inclusion of subdoxastic states. However, the requirement for *evidence* of reliability, that should be available on reflection, is epistemically value-laden, and so prevents the theory from fully satisfying the naturalistic commitments developed in Chapter 1.

formation, and so forth. It is at this level that all of the processing and work goes into actually generating the "seeming" that Barack will eventually have, and so this level plays a crucial role in the proper story of belief formation and updating.

While these events are sub-conscious, they already have many of the properties epistemologists have traditionally expected to find in theories of justification and their contributing parts. For example, highly successful connectionist models of facial recognition processes (Cottrell, 1990,[73] Laakso and Cottrell, 2000, and Cottrell et al., 2014) make use of relationships that could be described as "evidential" in nature. When identifying whether a face is male or female, different portions of the visual data are compared to proto-typical male and female faces, and the data that is similar gets counted as "evidence" for categorizing the face similarly to the prototype it resembles. These low-level networks are also sensitive to "defeaters", where a process can be progressing toward a certain categorization or outcome, and then have the trajectory radically altered (or "defeated") by a separate input or module's output processing. Additionally, we can ask the same questions about the processing at this level as we are interested in for justification. How should the visual cortex operate? When and where did a blind person's visual system make a "mistake"? What could correct it?

It is not a coincidence that these types of questions can be asked. The sub-conscious, subdoxastic components of our cognitive lives are essential to answering the questions that an epistemic theory of justification aims to address. As such, not only are these low-level states legitimately available for use in the construction of new theories (like the one under development here), but their inclusion is essential if we are to achieve a successful and genuinely informative

[73] Also, Churchland, 2007, p.136-153, discusses the philosophical importance of Cottrell, 1990 at length, and is recommended to the interested reader.

theory which holds true to the naturalist's scientific commitments. Let us look at the case for their inclusion in more detail.

First of all, our scientific research continues to amass evidence showing that a tremendous amount of key brain processes are sub-conscious or function as "dark-processing". This is a serious problem for traditional epistemic theories like foundationalism or coherentism, since it shows that our brains deal in far more than just the beliefs, reasons, and other mental states with which epistemology has generally concerned itself. At the neural level, excitatory or inhibitory signals may be received from neural clusters dealing with the monitoring of hormonal levels, or with tracking the background noise that falls far beneath our conscious notice, or any other number of sub-conscious things. Even if they have not been consciously noticed by the subject, however, more and more psychological research is showing that subconscious phenomena do influence the agent in a decidedly epistemic way.[74] Whether one accepts Freudian-style repressed memories or not, there are other well-documented phenomena such as the framing effect, the availability effect and the anchoring effect. Advertisers have long known that just getting a consumer to read the words "Limit of 12!" on a sale placard often will work subconsciously to make the consumer accept and act upon a belief that he or she should buy more of the product than was otherwise intended.

One approach that offers some insight into some of these phenomena is Dual Process Theory, which has proven to be both popular and influential in psychology, cognitive science, and other fields. This commonly used approach draws a distinction between two different types or "systems" of cognitive processing. "System 1" is what is used to immediately answer "1+1=?" or to instinctually grasp that a nearby growling lion poses an imminent danger. System 1

[74] Nisbett and Wilson (1977) is a classic example of this kind of work, also see Kahneman (2011).

performs these tasks rapidly by drawing heavily on association, and so is fast, efficient, automatic, and intuitive, but not very flexible. “System 2”, on the other hand, is deliberative and slow, but very flexible, and is the type of processing engaged to solve a complex long-division problem or carefully plan the next move in a game of chess (Kahneman, 2011). This way of categorizing our cognitive processes has proven to be quite fruitful in psychology and other fields, and I think shows great potential for epistemology.

One problem this approach helps us see more clearly is that most traditional theories in epistemology, as well as many of the standard definitions and concepts employed, have been rooted in a focus on only the processes included in System 2. This has happened because, as Kahneman (2011, Ch. 2) points out, we all individually tend to identify our “self” with the processes included in System 2. It is the deliberative, analytical, focused, reasoning that occurs in System 2 that just seems like the real “us”. Epistemology has therefore focused on the beliefs and procedures associated with our consciously accessible systems, examining what deliberative, conscious processes we ought to use in our thinking about the world, while largely ignoring System 1. While philosophers have of course been aware of the importance of our basic sensory perception, without an adequate scientific understanding of these low-level processes, there has, until recently, been no other option but to just treat the outputs of System 1 processes such as object recognition, memory access, etc. as “given”, “basic”, “primitive”, “foundational”, etc. Because of this, epistemic theories have traditionally proceeded by defining justification as “internal” in the sense that epistemology *starts* with the outputs of System 1 and then studies what we consciously do with that data and what inferences we draw from it. However, it seems that this is no longer adequate given what our science has discovered about the important role that System 1 processes play.

I think System 1 activity, in addition to being essential for our success and mental life, clearly exhibits important, epistemically rich aspects. Driving a car on "autopilot" (perhaps while chatting with a passenger) takes in sensory information about the surroundings, categorizes them, and forms representations that are then used in the planning of complex behaviors to accomplish the goal of driving safely. Sometimes, the "evidence" is used properly by the system and sometimes it is not (and *that's* when accidents happen!) Even understanding a simple sentence such as "The cat is on the mat" in our native language involves complex procedures of memorial access to our stored information about the meanings of the component words, the syntax rules of the specific language, etc. These and many other System 1 processes seem to involve processes that are clearly epistemic in nature. But if we accept the role of System 1 in our epistemic undertakings, as I think we should, this immediately requires that the traditional understanding of the mind or agent that participates in epistemic practices needs to be redefined, and that our traditional requirements such as "internal accessibility" must also shift. It appears that any high-level cognitive system is influenced by myriad sub-systems, and so any correct epistemic view must take these sub-conscious processes seriously and consider them as potential components (J-factors) in theorizing about how we evaluate a belief's justification.

Additional reasons for including subconscious states in our epistemic theorizing are found in anomalous psychological cases like blindsight. Patients with blindsight are not *consciously* aware that they have any visual information available in the affected visual field, but it is there and available for use by other parts of their cognitive systems. One patient was convinced that no visual information was available, and yet was able to successfully navigate a room filled with obstacles (de Gelder et al., 2008). This patient's cognitive system created a complex series of representations, performed calculations, planned motor movements, and

updated spatial representations as needed in order to accomplish the goal of crossing the floor. Traditionally, philosophers would have denied that this patient had any beliefs such as "there is a chair directly in front of me", but I would say that this patient should at least have spatial "beliefs*" attributed to him or her, since the representations are used in so many of the ways that regular beliefs are. To deny these subconscious information states epistemic status simply because they cannot be introspected by the "mind's eye" of consciousness or verbally reported strikes me as unacceptably *ad hoc*. In blindsight, information is being received, processed, and acted upon by a set of interconnected modules, and this representation and processing can be more or less successful. Sometimes blindsight gets things right and sometimes it does not, just like regular sight. There appear to be many interesting questions to be asked about what constitutes the difference between "good blindsight" and "failed blindsight", when is there adequate evidence for "belief*" *P*, and what role do and should blindsight "defeaters*" play? These sound very much like appropriate epistemic questions to me, and suggest that epistemology should concern itself with providing a theory that has the resources to address these questions, even though they apply to phenomena inaccessible to the agent's consciousness.

Finally, while the nature of consciousness is far too large a topic to be directly engaged here, there is a considerable amount of work being done that seems to show that consciousness is not actually what it has traditionally been thought to be. In light of current empirical findings and theories, many neuroscientists, cognitive scientists, and naturalist philosophers completely reject the notion of a "mind's eye" or "Cartesian theatre" that observes our thoughts, perceptions, and memories. This shift has played a significant role in causing an increasing number of philosophers to reject internalist theories. For example, Kornblith has argued that

> BonJour's internalism, and, indeed, all other internalisms, are motivated by a Cartesian view of an agent's access to her own mental states. This Cartesian view is… untenable, and, accordingly, so is internalism. (Kornblith, 1988, p. 1)

There has also been a related recent deflationary movement that suggests that there is not actually any one system in the brain that corresponds to "consciousness"; instead 'consciousness' is argued to merely be a word that we use to talk about multiple, disparate cognitive systems (Dennett, 1991). If there is no such thing as "consciousness", it would obviously be a mistake indeed to have it assigned such a key role in our epistemic theories!

The findings of this recent work seem to give us very good reason to think that we should not automatically privilege conscious states so heavily in our epistemic definitions and theorizing. Indeed, even before scientific research started casting doubt on the existence of consciousness, using conscious access for delineating relevant mental states was fraught with difficulty. Thoughts and perceptions can enter consciousness, and then suddenly go out of it, with no explanation available for how this happens. Consciousness has always been so poorly understood that it should be no surprise that it has caused many internalists to have to grapple with thorny topics like the differing roles of occurrent and non-occurrent thoughts, what counts as "accessible", and what it means to specify that access must be "reflective". The combination of longstanding difficulties with employing consciousness as a central epistemic criterion and recent scientific findings that cast doubt on the very legitimacy of the concept should be more than adequate to motivate the shift that I have in mind.

If we reject conscious access as our criterion for differentiating between what we consider internal and external epistemic factors, however, a different criterion will clearly need to serve this role. To this end, I now assume that the internalism and externalism debate should be recast in terms of whether justifying factors are internal or external to a *cognitive system*. This approach will permit a more scientific account of delineation, and will recognize the

importance of subdoxastic states to the justification of our beliefs. It may be that my proposed use of the internalist label seems to be too wild a departure from its traditional meaning. However, I certainly think that the demarcation that I have illustrated is a legitimate and potentially fruitful one. So, if this redefining of "internalism" seems unacceptable for some reason, please feel free to read the remainder of my project as advocating a theory of justification which could be labeled "internalist*".

3.3 Pragmatism / Pragmatically Constrained

As an internalist theory, endo-reliabilism serves to decouple the reliability of the belief-forming processes from the truth of the beliefs that they form, and instead makes it a function of their holistic coherence and conditional probability relations. This may be unsettling to some readers, as the attainment of true beliefs is generally seen as a fundamental epistemic goal of a rational agent. As unsettling as it may be, I think that the arguments in chapter 2 demand that we summon courage and confront the paucity of our epistemic situation head on. As finite creatures, inherently restricted to our epistemic perspective and subjective experience, the coherence of our "web of beliefs" (or, better, given the inclusion of subdoxastic states: our "cognitive web"), perhaps additionally constrained by the other theoretical virtues of simplicity, fecundity, explanatory power, conservatism, etc., is as far as the grounding for our beliefs *can* go.[75] Our sensory experience provides unavoidable,[76] raw, organization-requiring, input into the cognitive

[75] See Quine and Ullian, 1970, for more discussion of the theoretical virtues and their roles. Some authors, however, such as Bas Van Fraassen, are skeptical of the role that theoretical virtues should play in our reasoning. See his *Laws and Symmetry* (1989) for discussion.

[76] The fact that we cannot cognitively disable the stream of sensory input allows the avoidance of a worry that notes that a "web" comprised solely of *any* single belief is optimally coherent. BonJour (1985) addressed this objection, but I think it simpler to just note that we, by virtue of our neurological systems' functioning, have sensory perceptions across time, and until they are subsumed under *some* organizational

web, and the web natively incorporates desires, such as to "avoid pain" and "seek pleasure". This means that what our epistemic situation *does do* is provide the requisite resources for our cognitive webs to grow and optimize in ways conducive to the satisfaction of our pragmatic desires and goals.

If one allows that "truth" is itself a theoretical term within our cognitive web, and that we find ourselves believing things such as that *coherence is a guide to truth* and that *we are not "spinning frictionlessly in the void",*[77] it seems that we can tell an adequate story about how things such as sense experience, evidence, justification, and truth are all interrelated *internally*, and thus provide ample resources for explaining both our uses of these concepts, *and* the (mistaken) intuition that truth is something other than a part of this internal web.[78] While my approach will likely not be satisfying to many more traditionally-minded epistemologists, the commitments of the naturalist approach ought to make many naturalists feel satisfied with a theory based in acknowledging the self-policing nature of our cognitive systems, and that it adequately explains the successes and failures of our cognitive system in attaining our pragmatic goals.

The move suggested here is not without precedent. Pollock and Cruz think that, "practical and epistemic cognition are evaluated as a package. The ultimate objective is not truth, but practical success through the operation of epistemic norms" (Pollock and Cruz, 2004, p. 27).

---

structure, they yield rampant incoherence (for example, just one rapid visual experience can yield "redness and not-redness"). Just by *organizing* these sense data, coherence and the other theoretical virtues immediately become centrally relevant.

[77] as John McDowell refers to a coherent web of beliefs that cannot be known to make contact with objective reality, in his *Mind and World* (1994).

[78] We also view some coherence relations as *better* than others, not because they are more truth conducive, but because they are more pragmatically useful to us. This will be explored more fully in the next chapter.

While we may hope for some pure and optimal grounding for our epistemic work, we may end up simply having to accept the limits of our situation and realize that a pragmatic grounding is the best we are going to get. But after all, those of us attracted to the naturalist approach are generally willing to jettison our preconceived notions, hopes, and expectations for a theory if the "experimental" data require it.

It seems to me that our traditional notion of epistemology as oriented toward truth is the next piece that must go. But what might this look like? Let us step back a moment and ask ourselves: Why is the attainment of true beliefs seen by so many epistemologists to be a fundamental epistemic goal of a rational agent? Since naturalists are not likely to offer a response that asserts that the intrinsic value of truth is a brute, primitive fact knowable *a priori* or that true belief is "good for the soul" (or something similar), we can anticipate many answers that are variations on the theme that "true beliefs will better allow an organism to survive, reproduce, and attain its goals in its interactions with the surrounding environment".[79] This answer, however, seems to relegate the value of truth to being purely *instrumental* in the attainment of the organism's pragmatic aims, since it is the organism's survival, reproduction, etc., that are valuable in themselves to the organism, and true beliefs are presumed to be more conducive to these ends than false beliefs. If this is the case, then it seems that it is not *truth* that is valuable, but rather *success* in attaining whatever goals an organism has. (The move I am making here will no doubt seem familiar, since it is similar to the well-known arguments by Larry Laudan [1981] and Bas van Fraassen [1989] we examined previously concerning science and its relation to truth.) Scientific practice, and, I am claiming, general epistemic practice, can

[79] For example, Kornblith (2002) strongly advocates for a view along these lines, as do Bishop and Trout (2005).

and should be seen as concerned *only* with solving various problems and improving predictive, explanatory, and technological success.

Let us consider an example. If one imagines seeing a physician seeking help with some unfortunate medical condition or other, most people will initially say that they want the physician to be prescribing medicine or other procedures based on a correct, and *true*, understanding of the condition. However, if we look more carefully at what is going on in this case, I think it will eventually become apparent that neither the physician nor the patient cares about whether the physician's beliefs about the actual cause of the condition and its ideal treatment are *true*; what matters is whether the physician's beliefs and practices will lead to success in treating the condition. If we feel hesitation in agreeing with this, I think that it is likely that we are thinking that a false belief will *eventually* spell trouble, perhaps with another patient or in preventing the discovery of an even more effective cure, etc. It is important to note, however, that these failings that we think of as resulting from a failure of the beliefs in question to be true, are also easily (and I think more accurately) seen as failures of the beliefs to be success-conducive.

At this point, one may object by saying that it is a belief's truth that makes it success-conducive, and so my above arguments in favor of pragmatic success as the measure of epistemic norms are implicitly asserting truth's importance to us. It seems, however, that there are situations in which a certain false set of beliefs is more advantageous in attaining pragmatic goals than a set of beliefs that exactly corresponds to the objective state of affairs. If we imagine a set of beliefs that is adequate *in every way* for attaining all of the associated pragmatic goals, but is strictly *false*, perhaps because it categorizes phenomena in inaccurate ways, but ways that

make the system of beliefs easier to work with, it seems that we are actually better off with the false set of beliefs because it offers us advantages in attaining our goals.

As a classic example of this goes, if one's only interest is to build a bridge, one is far better off with the false theory of Newtonian mechanics, than with quantum mechanics (which seems to be at least closer to being true). On a related note, one of Goldman's standard examples in the process reliabilism literature is a discussion of forming a belief that one is seeing a mountain goat. And as an externalist, Goldman holds that the justification is tied to the ratio of true beliefs to false ones that the relevant process generates. However, our best scientific theories at the moment tell us that every time we perceive a mountain goat or any other "solid" object or surface, we are "seeing" something that is ultimately inaccurate.[80] It is *useful* for us to interpret enormous complex masses of electrons, protons, neutrons, quarks, Higgs Bosons, and so forth in an inaccurate but very useful way, using (artificial?) categories that help us attain our goals. Whether animal watching, hiking, or constructing a bridge, it is not always desirable to have our beliefs oriented toward the truth.

There have also been a number of recent studies that show that people who are more optimistic than the truth of their situation supports actually fare better in many ways than their counterparts who proportion their expectations more closely to reality. One of the most shocking of these studies is Reed et al. (1994), which shows that AIDS-positive patients who exhibited "realistic acceptance" of their situation have increased mortality, while patients with unrealistic optimism live longer. Research by Taylor et al. (1992) also provided evidence of the extensive psychological benefits conferred by what could be seen as "excessive optimism". If this is

[80] Depending on one's preferred theory of semantic content, this may or may not be inaccurate, so perhaps there is a way around this. However, I take the more general point about useful interpretations and categorizations to still be applicable.

possible, that a false set of beliefs could ultimately be more success-conducive for an epistemic agent than a true set, then it seems that the connection between truth and success is not necessary, and we have only to ask ourselves which one we would think *should* be adopted to see which of these is actually of epistemic importance to us.

A further interesting argument against the importance of the truth of our beliefs is provided by Stephen Stich in his book, *The Fragmentation of Reason.*[81] A fairly standard naturalist view of the truth-evaluation of beliefs holds that a *belief* is a brain state that somehow maps to a particular proposition.[82] On this account, a *true belief* is a brain state that maps to a *true* proposition. Stich, however, points out that there will always be more than one way to *map* brain states onto propositions,[83] and so there will also be multiple ways of mapping different sets of beliefs onto any one set of propositions, including the set of *true* propositions. Even presuming a unique set of true propositions exists, there will be a set of beliefs, *beliefs**, that maps to the set of true propositions in one way, and so are *true**, and a second set of beliefs, *beliefs***, that maps in a different way *onto the same set of true propositions*, and so are *true***, and so forth. The challenge to the externalist then, is: what makes having *true beliefs* preferable to having *true* beliefs* or *true** beliefs*? A non-*ad hoc* answer to this does not seem to be likely, and further, it seems conceivable that one set of beliefs, say, *true**** beliefs*, could even end up having certain advantages over the set of *true beliefs*— perhaps being more compact or simpler

[81] Stich, 1990, p. 101-127.

[82] Or, on other popular accounts: a truth-maker, a state of affairs, a set of possible worlds, etc.—these terms are freely interchangeable for 'proposition' in what follows.

[83] Unless one holds that there is a fixed, innate meaning for each brain state, it seems that there will be some way that mappings from brain states to propositions can be varied. For example, under a causal theory of reference, we need only imagine the causal conditions being altered in such a way that the physical brain state in question instead came to represent a different proposition.

in some way that makes it easier for a finite epistemic agent to utilize that set of beliefs in attaining goals.

These arguments could indeed leave one wondering if truth could be removed from our discussion of epistemic norms altogether. The theory I offer in the next chapter aims to provide a formal description of a set of belief-updating practices by an agent that seem to result in exactly the revisions that traditional epistemologists would claim that a "rational" agent should make. However, these practices and their formal description do not make any use of the notion of truth. In other words, it implements a belief-updating strategy that we recognize as squaring with most of our intuitions about belief justification, but that does not depend on a connection to truth. If it turns out that this is the belief-updating system that *we* use, then it is hard to see what causal-explanatory role is left for truth to be playing. Certainly, we use the term 'truth' in our discussion of our intuitions of how belief updating should work in response to evidence, but if the same updating can be captured by purely probabilistic models, then it seems that *truth* is just a "short-hand concept" that stands in for the probabilistic machinery that is doing all of the actual causal-explanatory work.[84] This also provides an error theory for why truth seems so central to our epistemic activities. Many epistemologists have a very strong intuition that truth is the aim of our epistemic efforts, but then, if the picture that I am painting in this section is correct, it would be quite natural to have this intuition even though it is inaccurate.

The above arguments provide reason for thinking that truth is not necessarily an important component of epistemic normativity, and so my theory's omission of a connection to

[84] While I will not explore it here, a parallel argument may be possible that indicates that even *justification* does not do any causal-explanatory work, and so can be eliminated from our epistemic discourse. This would be deeply upsetting to most epistemologists, but would simplify the epistemological project greatly by restricting our investigation to probability relations and their mechanisms.

truth is not damaging to the theory. Indeed, these arguments suggest that the tradition of largely constraining epistemological investigation to truth-centric theories has been, while understandable, somewhat misguided.

As we've seen above and previously in Section 2.3, some prominent philosophers have indeed worried about how we understand truth and whether it really is the aim of justification. Given this, it seems that it is worth investigating whether we can provide an account that fits our intuitions and does the work we expect of a satisfactory epistemic theory without needing truth as a component. This would, at the least, allow us to remain neutral about several difficult topics without thereby endangering the rest of the theory by tying its success to a particular view on truth and the role it plays.

3.4 Conclusion

At this point, the main *desiderata* that my theory aims to satisfy should be clear. I aim to provide a theory of justification that: satisfies naturalist commitments, retains the focus on causal history and process that helps make standard process reliabilism tempting, avoids the objections against externalist theories by confining itself to relata internal to the agent's cognitive system, respects and explores the rich and interesting roles subdoxastic states play, and is properly bounded by our epistemic situation and is therefore pragmatically-, rather than truth-, oriented.

Chapter 4: Endo-Reliabilism

This chapter develops a new naturalistic, internalist theory of justification, which I call "Endo-Reliabilism". The previous chapters have established several *desiderata* that I think should inform and constrain a theory of justification:

1. The theory should be naturalistic,
2. The theory should focus on process and causal history,
3. The theory should be internalist, as I have defined it, and only utilize relata internal to the cognitive system (in order to avoid many of the objections to externalism),
4. The theory should be genuinely informative about, and make use of, the subconscious/subdoxastic states that constitute a sizable and important share of human cognitive processing,
5. The theory should be pragmatically oriented and constrained,
6. And, if possible, the theory should have normative force or at least provide an error theory for why justification seems normative.

4.1 Modeling Reliability

To develop a theory of justification that meets the *desiderata* set out above, I suggest a model of reliability similar to that formalized by Bovens and Hartmann (2003), where we define reliability *endogenously*. While Bovens and Hartmann intend their model as a way to understand Bayesian commitments to different information sets delivered by, for example, witness testimony or scientific instruments, I will argue that this model provides the proper tools for

defining reliability in the same *endogenous* manner at the lower level of reliable *cognitive* processes. Before employing this model at the level of our cognitive processes and procedures, however, it will be helpful to describe the key elements of the proposal Bovens and Hartmann advance.[85]

Bovens and Hartmann discuss one example of an imaginary detective investigating a murder witnessed by multiple by-standers, as a way of gaining insight into how a responsible epistemic agent ought to evaluate a set of pieces of information (2003, p. 13). We imagine, first, the detective interviewing each bystander individually and writing down each proposition that the witness reports in a notebook. Each of the witnesses is presumed to be reliable to some partial degree (meaning that the witness is neither infallible in making reports of this nature, nor always mistaken, or more formally: $0 < r < 1$), and to make his or her reports based on direct experiences during the murder, so that the testimony can be considered independent of all other witness reports. The set of these propositions reported by an individual witness yields one information set *S.* As the detective accumulates information sets *S, S', S''* and so forth, these are all combined in the notebook into ***S***, which is an information set containing the different information sets reported by each witness. The detective then evaluates ***S*** to determine how confident she can be in the accuracy of the total description given.

For example, if we imagine all of the witnesses seeming equally trustworthy and competent, then *ceteris paribus*, an information set ***S*** = {[the culprit had a French accent], [the culprit was wearing Coco Chanel shoes], [the culprit drove off in a Renault]} should strike us as deserving more of our confidence than ***S**** = {[the culprit had a French accent], [the culprit was a

[85] Owing to the complexity of their formalization, the specific equations and formal Bayesian networks of their account cannot be included here. The interested reader is referred to the technical treatment found in their book, *Bayesian Epistemology* (2003).

male], [the culprit had dark hair]}. An important difference in these two sets of information is the degree of coherence between the information that they contain. If we imagine that all of the propositions in both ***S*** and ***S**** have equal prior probability (owing perhaps to some other background information that the detective has available), and that the witnesses making the reports have identical degrees of reliability, then ***S*** is to be preferred to ***S**** because the propositions in ***S*** are intuitively more coherent—they just "fit together" or "hang together" better. One way of formalizing this intuition is to examine how the probabilities of each proposition in a set would be affected by the truth of other propositions in that set. In other words, we are interested in the conditional probabilities of the propositions in each set. Considering the propositions in ***S***, the probability that the culprit had a French accent is made considerably higher if it turns out that indeed the culprit was wearing Coco Chanel shoes and did indeed drive off in a Renault. However, the probability that the culprit had a French accent receives a much smaller probabilistic boost (if any at all) if it turns out that the culprit was indeed male and did indeed have dark hair (as we find in ***S****).

Another illustrative example is Bovens and Hartmann's "Tweety" case (2003, p. 29). If we learn from independent sources that an acquaintance (1) has a pet bird named Tweety and (2) the pet is a ground dweller (and so cannot fly), this forms a fairly incoherent set (although not a fully contradictory one). Given that most types of birds can fly, learning the information in either (1) or (2) actually lowers the probability of the other piece of information being correct. However, if we were then to hear from another mutual friend that (3) Tweety is a penguin, the new information set of {(1), (2), and (3)} is now highly coherent. If it turns out the person indeed has a pet bird named Tweety, and Tweety is indeed a Penguin, then (2) is now highly likely to be accurate as well. (The other propositions in the set receive a similar probabilistic boost, *mutatis*

mutandis.) The addition of the information that Tweety is a penguin makes the other information "fit together" better, and increases the coherence. Note though, that the new information eliminates the possibility that Tweety is an Ostrich, Emu, etc., and so results in an information set that, while more coherent, is, strictly speaking, objectively *less* likely to be true since fewer possible ways the world might be will satisfy the statements that compose the information set.

Returning to the previous example, the detective's confidence in the truth of an information set ***S*** should, however, be informed by more than just how well the included propositions cohere with each other. The prior probability of the propositions that the set contains and the reliability of each of the witnesses reporting the propositions are also important factors. Holding any two of these three factors fixed and altering the third will directly affect the confidence one should have in the resulting information set.

While Bovens and Hartmann provide a detailed treatment throughout the book of all three of these factors, being especially interested in formal metrics for measuring coherence, it is the role of, and conceptual notions involved in, reliability that are our primary concern for developing our new theory of justification. For convenience in presenting their formal model of these three factors, Bovens and Hartmann initially treat the reliability of the witnesses as defined *exogenously*, being simply specified as a value to be input into the information set ranking specified (2003, p. 45). For example, perhaps the detective knows from experience that eye-witnesses tend to get facts right three-quarters of the time, and so she assigns all witnesses a reliability parameter $r = 0.75$. Or, perhaps the detective is more fine-grained in this assessment, assigning (female) witnesses wearing the nun's habit an r-value of 0.90, and motorcycle "enthusiasts" with extensive tattoos, eye patches, and apparent anti-law-enforcement attitudes an

r-value of 0.27. With these reliability parameters somehow *given* as inputs into the model, it is then possible to evaluate and compare different information sets.

4.2 An Endogenous Reconstruction

As Bovens and Hartmann recognize, however, "[w]e may not know beforehand whether the witness is reliable or not. So we construct a model with witness reliability—thus conceived—as an endogenous variable" (Bovens and Hartmann, 2003, p. 57). Essentially, what we are after is a way to adjust our assessment of the reliability of each individual witness based on factors internal to the model without having a direct estimate of the witness's reliability on the particular report in question available to us. We want to use these factors to determine where the witness falls in reliability between being infallible (r = 1.0) and giving reports that are no better than random chance (what Bovens and Hartmann term a: r="randomization parameter" akin to flipping a coin, or rolling a die and reporting according to the result).

In order to make this assessment, numerous relationships between the reports and witnesses have to be considered, so let us return to our detective example, and examine some of them. Imagine the detective has interviewed the first witness, who reports "I saw the butler do it". If the second independent witness reports seeing the same thing, the detective will tend to consider the reliability of both witnesses increased by the corroboration of their reports. While this increase is not linear, there is a generalization that after n reports of the same thing, a matching report from the $n + 1$ witness increases both the reliability of the $n + 1$ witness and all n previous witnesses. As Bovens and Hartmann observe, "[w]e may initially distrust a set of sources, but if they provide us with the same (or highly coherent) information, our confidence in their reliability is increased" (Bovens and Hartmann, 2003, p. 56). Similarly, a report from a witness dissenting from the reports so far (for example: "I was with the butler at the time of the

murder so I know it was not him") requires an epistemically responsible agent to adjust negatively the reliability assessment of *at least* one of the witnesses. Which witnesses are negatively affected by this, and to what degree, will depend on the specific numbers involved. Having the second witness dissent in this way will impact the reliability assessment in a different way than having 99 witnesses agree, and then having the 100^{th} witness dissent. (In the latter case, we tend to only downgrade the reliability of the 100^{th} witness, whereas in the former case, we may conclude that both witnesses should be treated as equally unreliable until additional information can be found to tell us which account to trust more.)

This relationship can be expanded to the more generalized notions of positive or negative relevance between the reports in such a way that partially matching or conflicting reports can be accommodated by the model (for example, reports such as "the murderer was a tall male" and "the murderer was a male"). This relevance relationship also influences perceived reliability after numerous reports by the same witness. We can easily imagine the detective's confidence in $witness_1$ being shaken by the witness' changing her story the second time she is asked what happened. The role this plays in larger numbers of reports by the same "witness" is easiest to envision if we instead consider a scientific instrument, say a particular electrical current detector, that generally yields the same report in the same experimental conditions but every once in a while yields conflicting results. The number of cohering reports will obviously influence whether we think that the reliability of the instrument is suspect, or whether we turn our attention to other components in the experimental setup to find an explanation for the deviating report.

Besides the *number* of reports and the *relevance relations* between them, the *plausibility* (and more precisely, the *subjective prior probability*) of the individual reports the witness provides also affects the endogenous reliability measure. A witness at a session of the *American*

Philosophical Association reporting that the culprit was wearing a tweed jacket should, *ceteris paribus*, be considered by the detective to be more reliable than a witness making the same report on a tropical beach. As Bovens and Hartmann observe, in addition to our probabilistic assessment of the particular report (or hypothesis) *actually* being the case, we also consider how likely we think a particular witness (or instrument) is to make a certain report.

> If we are dubious that the hypothesis is true, then we tend to blame a positive report on the lack of reliability of the witness—and even more so if we know that an unreliable witness would be randomizing with a high probability of providing a positive report…[however, i]f we are confident that the hypothesis is true, then we take a positive report to be a reason to increase our trust in the reliability of the witness [and especially if we think that a positive report is unlikely.] (Bovens and Hartmann, 2003, p. 60)

Thus, the reliability assessment of a particular source of information is influenced by the interplay of all of these factors, which are themselves *internal* to the agent evaluating the reliability. When these various factors come together in the right way, the witness or instrument in question is seen by the epistemic agent as highly reliable, and the reports that have been provided have their subjective probabilities adjusted accordingly, which are then used by the epistemic agent in pursuing other tasks, such as weighing coherence with other beliefs, deducing entailments, guiding actions, and so on.

There are two final things to note here. The first is how closely connected this endogenously defined reliability is with the conditional probability relations between the reports. A reliable witness is identified as reliable because the reports offered stand in the right kind of probabilistic relationships to the other witnesses, reports, and prior probabilities involved; reports made by reliable witnesses have high-conditional probabilities as a function of this interrelation. Second, it is important to notice the important role that the *goals* of the detective (or, in the other case, the scientist utilizing the instrument) play. In the above discussion of the detective, we automatically assume that, by virtue of being labeled a "detective", the agent has the goal of successfully determining the identity of, and eventually apprehending, the criminal in question.

Presumably, if effective, the detective in our story is constantly reevaluating the witnesses, methods, and processes that he or she utilizes in relation to how successful they have been in accomplishing the goals that he or she has. A different goal or set of goals could potentially have a profound effect on many different parts of the account given so far. For example, if the detective is of the variety sometimes portrayed in film as lazy, jaded, and perhaps near retirement from the police force, this individual's goal might instead be simply to put in as little effort as possible. This could result in a willingness to accept any witness testimony as correct with little scrutiny (by assigning a high prior probability to *every* statement), or perhaps a goal to collect and examine only witness statements that are, say, two sentences or less in length. The situation becomes even more unusual if there were a detective who (for some strange reason that only philosophers would construct) has the sole goal of *failing* to catch the murderer. The takeaway from these (hopefully!) unusual cases is that the agent's goals are also an important part of the assessment story, and help to regulate the entire process.

4.3 Process Endo-Reliabilism

Having seen how this notion of endogenous reliability operates at what could be called the "agent level", let us now apply this same model to our cognitive processes. I contend that traditional epistemological process reliabilists can be seen as advocating a theory of justification that defines reliability *exogenously*, where the reliability of our various belief-forming processes and procedures are simply *given*, defined externally to the epistemic situation in which we find ourselves.[86] This is chiefly where standard process reliabilism deviates from internalist

[86] Interestingly, Audi discusses testimony by individuals with varying degrees of reliability as a motivation for process reliabilism (Audi, 2002, p. 227). But here, as is typical in these discussions,

leanings—it presumes that reliability is to be defined exogenously to the agent (and the agent's cognitive system), as a relation involving the world beyond (and thus not directly accessible to) the epistemic agent. Since it seems that, if we take the epistemic limitations of our subjective, human condition seriously, we must admit that any of our cognitive processes could actually be misleading us (but in such a way that we do not recognize their unreliability), the situation seems to push us to redefine the reliability of our epistemic processes endogenously. This change will be especially important if we wish to have justification available to function in a regulatory capacity within individual epistemic agents.

Where the account offered by Bovens and Hartmann dealt with "witnesses" thought of as either humans or scientific instruments, and dealt with assessing the reliability of their reports, the same relationships and modeling procedures can be applied to our internal cognitive processes—thinking of our various cognitive processes as "witnesses" and their processed outputs as "testimonies" or "reports". Following the move Bovens and Hartmann make, just as we can assess witness reliability across many reports by gauging coherence with other reports (endogenously), so we can with cognitive processes. The feedback and interaction from different levels and modules of our cognitive processing affects our assessment of the reliability of these processes, and it seems plausible that it is actually *this* "endo-reliability" of the involved cognitive processes that determines whether a resulting belief is justified. At this point, an illustration may prove helpful.

---

reliability is treated as exogenous. While convenient, treating reliability in this manner just does not seem to fit with the reality of our epistemic situation for the reasons previously discussed.

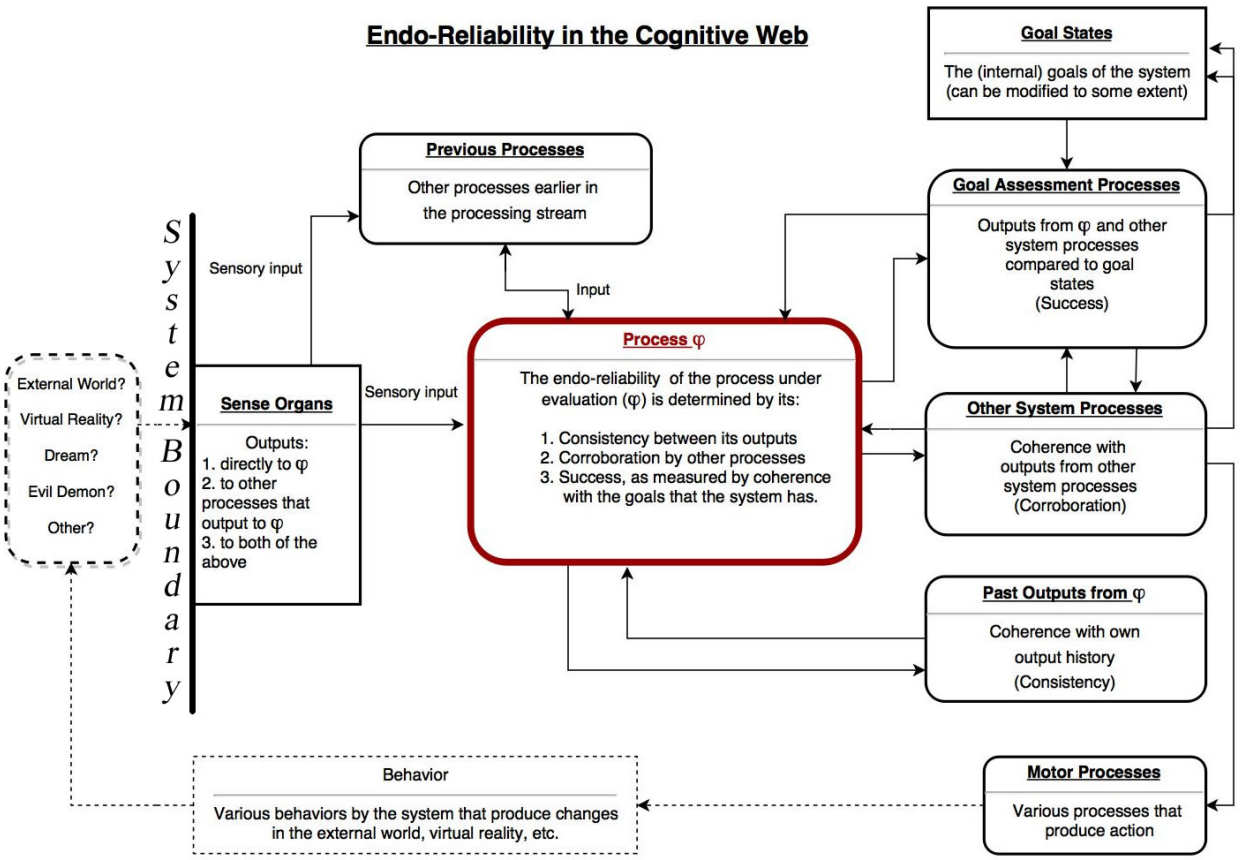

Figure 2 Endo-reliability in the cognitive web.

In Figure 2 above we see the various interrelations that determine the endo-reliability of a particular cognitive process φ. Depending on what type of process φ is, it may receive input directly from a sense organ, or it may be further removed and so receive input from other processes, or it is possible that φ receives a combination of input from the sense organs and other intermediary processes.[87] Wherever φ is located in the processing chain, its endo-reliability will be determined by its track record of producing consistent and coherent outputs, producing outputs that cohere well with the rest of the processes in the cognitive system, and producing

[87] Of course, if the correct hypothesis about the external world is actually that of a virtual reality simulated for a disembodied brain in a vat, then there will not be sense organs included in the story. However, the simulated input into the rest of the cognitive system will presumably remain.

outputs that either directly, or via the mediation or inclusion of other cognitive processes, produce states that cohere well with the goal states of the cognitive system. Concerning the last of these, the goal states, it is important to remember that it is not the satisfaction of the goal in the external world (say physically eating food) that is relevant here; instead it will be the attainment of the attending cognitive states (roughly, the *experience* or *representation* of eating food). In order to be used by the system for decision-making and assessment, its goals must be internal to it.

While it may not be immediately obvious in the above model, both subjective prior probabilities and epistemic defeaters are also incorporated. Let us examine each of these individually, before turning to some examples of the theory at work. Considering subjective prior probabilities first, note that a "prior" will be a pre-existing belief about how probable something is. As a pre-existing belief, this means that it was previously produced and maintained by one of the system's processes and must be stored in the cognitive system. As a result, when the endo-reliability of φ is evaluated by checking the coherence of its outputs with the outputs of other system processes, the prior will be among the things checked, and thus plays a role in the endogenous-reliability assessment. This is important because we saw above in the detective case that the endogenous approach to reliability includes the idea that the reliability attributed to a witness is increased or decreased in response to how well their reports match the priors of the detective. At the process level, the same effect can be seen in the change in endo-reliability that results from the degree of coherence of the process' outputs with the priors produced and stored elsewhere in the cognitive system.

Turning our attention to justificational defeaters, we find that these, too, are naturally and automatically built into the theory.[88] Let us look at why this is the case. I think that, at their core, defeaters across most (or all) theories of justification are ultimately about coherence relationships. Even if someone is a foundationalist, holding, for example, that justification results from proper support by a chain of good inferences grounded in privileged, basic beliefs (or sense data, etc.), the idea of that justification's being defeated usually involves another belief or set of beliefs that does not cohere with the original belief. It is the incompatibility, and lack of coherence between the beliefs, that does the defeating. And as we saw in Chapter 2, especially with the case of Norman the Clairvoyant, one of the main difficulties faced by early formulations of externalist process reliabilism was the difficulty in explaining how a belief (such as that there is no such thing as clairvoyance) could defeat the justification conferred by a reliable process. The difficulty here is that traditional process reliabilism does not include coherence relationships among the factors that determine justification.[89] It ignores most of the coherence relations and then "cherry-picks" certain incoherence relations for inclusion, without providing a principled

[88] There is a distinction that must be drawn between what Steup (1996, p. 14) calls "factual defeaters" (also sometimes called "propositional defeaters") and "justificational defeaters" (or alternatively, "mental state defeaters"). A factual or propositional defeater is a fact or true proposition that is external to the agent that defeats the agent's ability to have knowledge about something. If I get lucky and form a justified true belief on the basis of glancing at a previously accurate, but now broken, clock, at just the right time of day, my *knowing* the time is factually defeated by the fact that, unbeknownst to me, the clock is broken. If I were to discover that the clock was broken, my justification in the belief would then be defeated. On the other hand, a justificational or mental state defeater is a defeater internal to the agent, such as a conflicting belief or piece of evidence, that removes (or reduces) the justification from the belief or mental state in question. It is this type of defeater at work when my otherwise-justified belief that the paper in front of me is red is defeated by my co-occurring belief that the room is bathed in red light. It is this latter type that is relevant for my theory here.

[89] I think it is telling that Goldman (2011), Comesaña (2010), and other defenders of reliabilism have recently been trying to develop a hybrid approach that incorporates evidentialism into reliabilism, since a successful blending could allow *evidence* of the unreliability of a process to remove or prevent the justification of the resulting belief. See also Steup (2004).

account for why those coherence relations matter while others do not. As a result, many of the proposed ways of incorporating defeaters have a decidedly *ad hoc* feel to them, since they are not core components of the theory developed, and seem to have been tacked on to solve this specific problem. On the other hand, traditional doxastic coherentist theories of justification, such as that developed by BonJour (1985) and Lehrer (1990), often have the advantage that a theory of defeaters is "had for free" in virtue of the core commitments of the theory. If the justification of a belief is the result of its coherence with other beliefs (either the holistic set of all the agent's beliefs, or a subset restricted in some fashion), then a justificational defeater will be any belief or set of beliefs that, by failing to cohere properly, removes, reduces, or prevents the level of justification that the original belief would otherwise have had. This is an advantage that endo-reliabilism also inherits from the central role that coherence relationships play.

In endo-reliabilism, there are two ways that justificational defeaters manifest. First, since the endo-reliability of a process is a result of the coherence relationships that its outputs stand in with the rest of the cognitive system, anything that lowers the endo-reliability of a process defeats (or undermines) the justification that the process confers. Returning to Norman's case, if his belief that there is no such thing as clairvoyance (or, at least, that he is not clairvoyant) is justified, and so results from an endo-reliable process, this belief and the coherence relationships that justify it will serve to consistently prevent Norman's clairvoyance process from attaining the level of endo-reliability need to confer justification. And if his clairvoyance process did somehow become sufficiently endo-reliable to confer justification (perhaps as a result of new experiences or new information), it would necessarily involve a corresponding decrease in the level of justification enjoyed by the belief that he is not clairvoyant.

The second way that defeaters are incorporated in the theory occurs when an endo-reliable process generates an output, that would otherwise be justified, but which has that justification defeated (or rebutted) by the output from another process. It may be helpful to think of these outputs as "proto-beliefs", not yet having risen to the status of "full belief" (or all-things-considered belief) in the cognitive system. Consider again the standard case of a room bathed in red light. Under these conditions, if I glance at a piece of paper on the table, my visual processes will form the proto-belief that the paper is red. Presumably, my visual processes are endo-reliable, since they have a good track record of producing outputs that stand in the right probabilistic relationships with each other and the rest of my cognitive system, and so that proto-belief is justified. Imagine that I also have a proto-belief that the light in the room is red, itself resulting from an endo-reliable process, and so justified. In this case, whether the full belief "The paper in front of me is red" is adopted will depend on processing that occurs further "downstream" in the cognitive system, and the justificatory status of that belief, if adopted, will depend on the endo-reliability of the process used. To the extent that that downstream process is endo-reliable, it will have a history of making good "judgments" in cases like this one and having its outputs cohere well with the rest of the cognitive system, the information it processes, and its goals. Given the importance of the concept of defeat in epistemology, I think its deep and fundamental integration into the basic components of endo-reliabilism constitutes a considerable advantage of the theory.

With all of this in mind, let us define endo-reliability as follows:

> **(Belief) Endo-Reliability**: A cognitive process is endo-reliable to the extent that it tends to produce and sustain beliefs that have their probabilities increased conditional on the prior outputs of that and other processes, other items in the

> cognitive web and increase the probabilities of the prior outputs of that and other processes when conditionalized upon. [90]

This yields the following formulation as a new theory of justification to be considered:

> **(Belief) Endo-Reliabilism:** *S* is justified in believing that *P* at time *T* to the extent that the process of belief formation from which *P* results is endo-reliable.

To understand better the application of this idea, let us consider some examples. These are not meant to be complete descriptions or explanations, merely helpful illustrations of the theory described so far. And, given the naturalistic commitments I argued for earlier, I think it clear that our views on how our cognitive systems engage in these tasks should ultimately be settled by neuroscientists and others conducting empirical experiments, not by a philosopher writing from the "arm chair". Still, examining some sketches of how the theory of endo-reliabilism might apply will provide useful illustration of both the theory and its various applications.

4.3.1 Perception

Let us begin by applying endo-reliabilism to everyday sense perception. If we consider what justifies a visual perception of an object such as a kitchen table (in normal lighting conditions), we can evaluate the reliability of our visual processes endogenously. First, we note that our visual perception yields consistent perceptions of the table if we continue staring at it. Then we compare our visual perceptions with the reports of our tactile sensations when we reach out to touch it, as well as to our memories of buying and positioning the table in the room, and so forth, and we find that all of these "reports" from our cognitive processes are highly positively

[90] At present, this definition deals only with beliefs and not with other mental states. While traditionally epistemology has considered justification to be a property of beliefs, the approach that I am advocating here will shortly be extended to include other mental states as well.

relevant to, and coherent with, each other. Then we note that these visual perceptions have high prior probability, since kitchens are the types of rooms that frequently have tables, and so forth. All of these factors serve to establish the endo-reliability of our visual processes in these conditions. If, on the other hand, we were to see the same table floating in the sky, and find our tactile reports to be that of empty air, we would rightly come to doubt the reliability of our visual processes in making reports in these conditions (perhaps influenced by poor lighting conditions or hallucinatory drugs), and so would be unjustified in forming the belief that there is indeed a table before us.

It is not just the coherence between our senses that matter, however. The reliability assessment metric also ranges over the fit of the sensory information and the background conceptual framework that is in place. Imagine that you are in the deep woods, and see a fox scampering between the trees. You see it running, can hear the soft rustling of its feet, and, if you are skilled (and brave) enough, could touch its soft fur. Suddenly, it notices your presence, turns to you, opens its mouth wide and emits… the mechanical sound of a car alarm! Even though the perceived location of the sound and the visual perception match up nicely, the clashing between the experience of a fox making a mechanical car alarm beeping and the conceptual framework that leads to the co-activation of *FOX* and *NON-MECHANICAL* should certainly cast doubt on the reliability of our senses in this situation.

4.3.2 Memory

Imagine someone asks Smith what year World War II ended. Smith pauses for a brief moment and then answers “World War II ended in 1945.” Unlike in standard process reliabilism, where the justification of his belief depends on the past history of how many times the relevant memory process has gotten things correct (by remembering *truly*) and how many times it has

failed (and the ratio between them), endo-reliabilism will say the relevant process is reliable to the extent that the process has produced remembrances that, for example, cohere well with his other propositional memories (that WWII was still happening in 1944, that it was over in 1946), his visual memories of what he has seen in WWII photos and film footage (the types of clothing soldiers wore; the popular hair styles; the level of technological sophistication of the trucks, planes, and guns) and how these fit in with his conceptual framework of what was common in which time period, etc. It is also benefited by the fact that every time he thinks about the end of WWII, his memory produces the same year as an answer. He also remembers doing quite well on his high school world history exams, which helped him to attain his goal of getting good grades in school, and thus reinforced his belief that he has a good memory, at least for highly significant world events. As a result of the probabilistic relationships between these cognitive states and others, the particular memory process Smith used to answer the question he was asked has a good probabilistic track record, is endo-reliable, and so confers justificatory status onto his belief that WWII ended in 1945.

Compare this situation to the case of Jones, who happens to give the same correct answer, but whose memory process has a different degree of endo-reliability because he remembers failing almost all of his history exams in school; thinks (mistakenly) that in the 1940s, plastics were in widespread use in all vehicle construction and is perplexed about why none of the WWII photos he has seen include vehicles of this type, wonders why WWII radio operators did not just use their cellular phones to communicate, remembers being laughed at by his academic peers for his tendency to get historical facts wrong and therefore believes that he often fails in his goal to be respected and accepted by his colleagues. According to endo-reliabilism, even though both Smith and Jones give the same correct answer (and we could even imagine a situation where,

perhaps by Jones being incredibly and consistently lucky, their relevant memory processes have the same truth track-record and ratio of true-to-false beliefs), the difference in how well the outputs of their memory processes cohere with their other beliefs and cognitive processing outputs, makes a difference in their degree of justification.

4.3.3 Introspection

The next day, Smith is introspecting that he is in pain. Specifically, that his right hand is in pain. The justification for his belief, according to endo-reliabilism, is again to be found in the endo-reliability of the introspective process that he is employing. In this case, Smith's cognitive system has also produced a visual experience that his right hand is currently located on a stove, and that the stovetop dial is set to "On" with an accompanying indicator light. Smith remembers his mother telling him that hot stoves will burn him and cause him to be in a mental state of pain.[91] He remembers walking in to the kitchen, and remembers turning on the stove to cook dinner. Additionally, Smith believes that when he broke his leg last year (jumping from a run-away trolley in yet another philosophical thought experiment), he introspected pain in a similar way, was able to see the damage to his leg with his eyes, and heard several doctors tell him that the x-rays showed his leg was indeed broken.

What justifies Smith's belief? Endo-reliabilism will point to the *probabilistic* relationships between the outputs of the introspective process and the outputs of the other parts of Smith's cognitive system. Many philosopher's, however, take introspection to have a special

[91] Smith, as the philosopher's standard thought-experiment guinea pig of choice, may well have had a philosophically trained mother who spoke to him in *exactly* this fashion.

epistemic status, since it seems to them to be infallible, or at least highly unlikely to be wrong.[92] If introspection is indeed special in one of these ways, an endo-reliabilist will say that it is special precisely because of the *particular* probabilistic relationships its outputs enter into with the rest of the cognitive system. As a very quick and rough example, it may be that a cognitive system has a very central, fundamental goal to always have access to a representation of its own current status, no matter whether it is good or bad, because having this information is highly correlated with attaining states that cohere well with the system's other goals. This would mean that any output from an introspective process could receive a tremendous boost to its endo-reliability assessment because of how well *any* output about the system will cohere with the goal. The probabilistic nature of this account may well result in introspection falling short of infallibility, but I and others contend that this is actually the correct result. Well-known empirical work in psychology seems to decisively disprove claims to the infallibility or certainty of introspection, pointing to numerous cases where people, for example, are manipulated by experimental conditions but do not introspect the conditions having an effect on their actions, and instead introspect inaccurate explanations.[93]

4.3.4 Reason/Inference

Smith finds himself believing *P*. Smith also believes $P \rightarrow Q$, and proceeds to infer *Q*.

The cognitive process that produces the resultant belief in Smith stands in very strong relationships with many other parts of his cognitive system. Smith has had other processes

[92] There are many other interesting and complex philosophical questions about introspection, consciousness, and the relationship between the two, that are unfortunately outside the scope of the current project. For discussion of these issues, see Gertler (2011), Carruthers (2011), and Prinz (2012).

[93] For discussion of this empirical work, see Nisbett and Wilson (1977), Goldman (1986), and Kahnemen (2011).

produce highly coherent beliefs across a wide range of situations that are positively relevant to this inference. He has thought "if it is raining, then the ground is wet". He has seen it be raining, and then immediately seen that the ground was indeed wet. In fact, his memory tells him that he has seen countless instances of this inference at work, and when he has followed it, things have worked out well for attaining his goals. Additionally, Smith has never encountered a case where his cognitive process made this inference, and the inferred *Q* turned out to clash with the outputs of his sensory systems (appearing to be false) while the first two premises remained highly consistent with his sensory systems (and so appeared to be true).

Curiously though, Smith has some difficulty working with this same logical inference when it is abstract. Indeed, he was one of the many undergraduates who performed poorly on the famous Wason Selection Task.[94] Smith was presented with four cards, with numbers on one side and letters on the other. The cards he saw were showing: A, K, 3, 8. When Smith was asked what cards must be flipped over in order to see whether the rule "if a card has a vowel on it, then the reverse side has an even number" is true, Smith, like many of his peers, correctly flipped over the "A" card, but then unnecessarily flipped over the "8" card while failing to flip over the essential "3" card. Interestingly, Smith, also like many of his peers, had no such difficulty when he was asked to verify the rule "if someone is drinking beer, they are over 21 years of age", and then presented with the cards: drinking beer, drinking coke, 16 years old, 25 years old.[95]

In someone that is not trained in logic, the cognitive process used in working with the abstract or symbolic versions of the inference presumably does not enter into as many tight coherence relationships as one might hope. Perhaps the subject thinks that the inference kind of

[94] Wason, 1966.

[95] Cosmides and Tooby, 1992.

reminds them of something they saw in math class a while ago, or seems like one of those weird questions that psychologists (or experimental philosophers) are always looking for people to answer. But the (presumably) different cognitive process used to generate the subject's belief in the beer-drinking case is another story altogether. This process has strong connections to what the subject believes about alcohol, drinking laws, police enforcement and discipline more generally, and perhaps strong memories of trying to sneak into bars while underage, remembering stories from older siblings who were caught, etc. This furnishes this process with a much broader pool of probabilistic interrelations, and given the high coherence and positive relevance between them, accounts for the subject's higher degree of justification in this situation. This also means that endo-reliabilism would seem to have the interesting, and I think quite plausible, result that even a simple *modus ponens* inference is more justified for a trained logician than a layperson, because the logicians training allows the relevant cognitive processes to enter into many more, and far richer, interrelationships.

I should mention that I am aware that some more traditionally minded epistemologists will have a favorite analytic and/or *a priori* case for which they might think my theory unable to account. However, I do not find these kinds of objections compelling because I follow Quine in rejecting a distinction between analytic and synthetic statements, and, since any belief could be given up in the face of adequate empirical evidence, am suspicious of the existence of anything genuinely *a priori*. If I am wrong about this, and it were true that my theory cannot possibly tell a satisfactory story about these cases, that would of course be a problem. However, I have yet to encounter a compelling response to Quine's arguments offered in "Two Dogmas of Empiricism" (Quine, 1951), and it is by no means certain that my theory will be unable to account for all the cases that could be presented.

4.3.5 Testimony

Smith is sitting in class and the teacher says something surprising in lecture (perhaps that "The difference in time between when Tyrannosaurus Rex and Stegosaurus lived is greater than the difference in time between Tyrannosaurus Rex and right now.") While this initially seems unlikely to Smith, in that it does not cohere well with his other pre-existing background beliefs, he also remembers the teacher saying lots of other things earlier in the lecture that he did already believe. He remembers that teachers are generally well educated, and that they tend to either be experts in their material, or have (hopefully) done some research in preparing their lectures. His cognitive system has years of experience seeing others refer to schools as places where knowledge is transmitted, and thinks about his parents' words that he needs to get an education if he wants to be successful in life. He also looks around the room and perceives some students nodding in agreement, and others writing down what was just said in their notes.

Smith, however, has also learned to not take the testimony of the individual wearing the techni-color dreamcoat, living under the bridge, and shouting to himself, at face value. The last time he did this, the directions he asked for led him astray and prevented the attaining of his goals. The difference between these cases indicates that the credence we should give to any particular testimony depends on a great many factors relating to the reliability of the speaker or author. Endo-reliabilism is particularly well suited to incorporating these many factors (remember again the case of the detective discussed above).

4.3.6 Discussion

Explaining the reliability and justifying nature of the processes of perception, memory, introspection, reasoning, and testimony can be very difficult on other accounts, because it is difficult if not impossible to justify each one in isolation. To establish that any one of them is

reliable, one often has to utilize many or all of the other four. Endo-reliabilism, however, not only tolerates this interdependence, but encourages it because of the richer interrelationships that positively relevant outputs produce.

It is also worth noting that in the cases above and all others, endo-reliabilism will yield the same justificatory status for a belief held by a subject deceived by an evil demon as for a subject in a real external world, *ceteris paribus*. As argued in Chapter 2, this is the correct result that a satisfactory theory of justification must yield.

Hopefully, these examples have made the central ideas of endo-reliabilism clearer, and paint a picture of the applicability of the theory across a wide range of epistemically important situations. However, as I have pointed to previously, this discussion does not capture the actual epistemic "magic" because it is still at too high of a level of description.

4.4 Endo-Reliabilism at Subdoxastic Levels

While the above examples are meant to illustrate the theory under consideration, the justification for many of our beliefs will instead be found at much lower levels which may not correspond to things at our conscious level of experience. For example, the endo-reliability of a facial-recognition process may be a function of its probabilistic relationships to neural network functions that do not perform a function that is easily describable in everyday "folk psychological" terms. Still, the theory's probabilistic relationships and evaluation metrics suggested above will operate the same way at the neural level.

To get clearer on how this works, let us consider another example—one which operates at a subconscious, subdoxastic level, but to which we can still relate. Consider a visual process

responsible for detecting the edge of an object in a small part of a subject's visual field. If this process is endo-reliable, it will be because of factors such as the following:

a. the process continues to output the same (or highly coherent) output while the subject and object remain stationary;
b. when either the object or the subject moves (or saccades occur),[96] the process provides updated information that coheres well with the previous information (combined with motion-tracking reports from other cognitive processes);
c. the process outputs information that coheres well with the outputs of other processes responsible for edge detection in different, but nearby, areas of the visual field;
d. when the cognitive system has engaged motor processes to attempt to grasp, avoid, or otherwise interact with the object associated with the edge detected, the process provides information that coheres with the outputs of the relevant tactile processes, motor processes, etc.
e. the outputs of the process, and the relationships with other relevant processes, produce cognitive states that cohere well with the system's goals of grasping, avoiding or otherwise interacting with the object (in other words, the representation of successful attainment of the system's goals).

As we see in the above example, a subdoxastic process or neural network is endo-reliable to the extent that it tends to produce and sustain neural activations that stand in the right probabilistic relationships with the rest of the cognitive system's neurological processes.

[96] Saccades are rapid, unconscious movements of both eyes that occur in humans (and other animals) as part of the normal vision process (Bermudez, 2010, p.348).

Networks with high tendencies to do this will be weighted more heavily by the system through a neurological reconfiguration process. The exact manner by which neural networks and their synaptic connections are adjusted in the brain is still under investigation, but it seems likely that it heavily utilizes Hebbian Learning, where, as the saying goes, "neurons that fire together, wire together".[97] By adjusting the strength and "weights" of different neural synapses, making some connections more sensitive while inhibiting others based on their history of co-firing, the system retains information about past neural activity and can be said to have 'learned'. It is this raising or lowering of a synapse's activation threshold that corresponds to the trust or credence that a detective puts in a witness at the agent-level discussed above. Just like when a detective gets a report from a witness that has (so far) been evaluated as extremely reliable, she will often immediately pass on the report to a superior or act upon it in some other way, a neural network (once it has been sufficiently developed and trained) that receives an impulse or set of impulses through a synapse with a low activation threshold will readily pass on the impulse to the next neuron in the chain. A signal received over a synaptic connection that has been strongly inhibited is treated with the same hesitation that a detective would a known liar who wants to confess to the Kennedy assassination for the tenth time. So, just like a witness' reliability can be determined endogenously by evaluating the probabilistic relations of his or her reports with other reports from the same witness and others, as well as the background information available to the detective, the reliability of a cognitive process is a function of the fit of its activation outputs with the outputs of the rest of the cognitive system's components.

Some of the most important components, both in the human brain and in the theory that I am here advocating, are the neurological modules and processes responsible for assessing the

[97] See Churchland (1995, 2007, and 2012) for more discussion of how this might operate, as well as the current scientific evidence for it.

attainment of our myriad goal states (both large and small). Given the discussion last chapter that we are aimed at success (instead of truth), it is crucial that the system have the self-regulating capability to assess how reliably the current cognitive approach is working, and the feedback capabilities to make changes accordingly. Neuroscientific findings and standard connectionist approaches to neural modeling include the resources for telling a story about how this might work. Of course, neuroscientists and cognitive scientists still have many questions to answer and mysteries to solve, and as a naturalist, I will happily defer to the scientists' findings when they are made.

The story that I am about to tell mixes work done in artificial neural net modeling and findings and models from neuroscience. There are still very difficult and unanswered questions about how the brain implements, if it in fact does, the behaviors modeled by connectionist approaches to neural nets. But the artificial models used are still some of our best tools for modeling and learning about the function of neurons when assembled into networks, and so I will use some of these terms and techniques in what follows. My belief and hope, however, is that the general features, properties, and behaviors described here are reasonably accurate descriptions of what is going on at the much more complex and "messy" level of the biological brain.

In Figure 3 below, we have a fairly standard connectionist picture of a recurrent neural net, populated by models of neurons and their connections. The bottom layer represents neurons at the input layer, which pass information to the middle or "hidden" layer of neurons, which are in turn connected to the output layer of neurons. Each of the connections between neurons is made by a modeled synapse that has been configured with a certain weight or strength. Information processing proceeds both from the input layer towards the output layer and in the

opposite direction via neural feedback connections. For this example, however, we imagine that one of the input layer neurons (colored red in Figure 1) is receiving input from unreliable processes upstream. At first, these upstream "liar neurons" are treated equivalently to any other input, since without an externally specified, exogenous reliability value, the system has no training on the reliability of any other parts of the system.

Over time, however, the reliability assessment of the input from the liar neurons will be lowered, since either the lack of proper coherence of the outputs with the outputs from the other network segments will cause the signal from the liar neurons to be lost in the hidden layer, or else the feedback from whatever specialized module or process assesses the satisfaction of the system's goal states will eventually reconfigure the weights of the connections to weaken the connections from the liar neurons. In this way, over time, the unreliable network segment is identified by the system.[98]

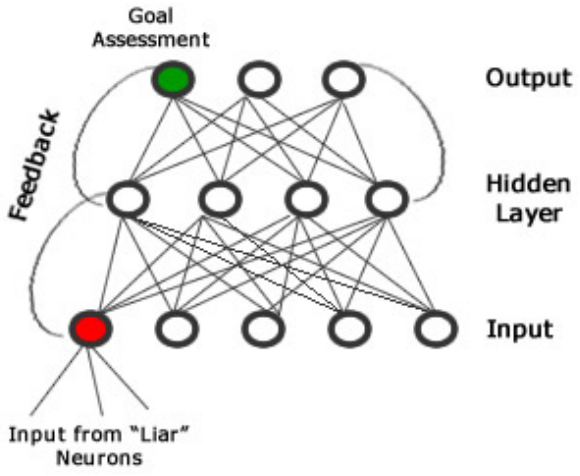

Figure 3 A sample recurrent neural net. (Not drawn to scale.)

[98] See Bermúdez (2010) and Churchland (1995, 2007, 2012) for additional information about backpropogation and other algorithms that can be used to train neural networks.

In artificial modeling, the goal assessment is often done by the experimenter, by comparing outputs from the net to the "correct" outputs associated with a set of training data, or by part of the network's learning algorithm. In human brains, it is possible that a similar function is being performed by other specialized neural networks that assess how well the process outputs cohere with the goal states of the cognitive system. Another possibility is that, instead of dedicated assessment modules, these functions are distributed throughout the network.

Of course, when something is not working, there is no automatic indicator of which part of the overall process is the problem, and so we find ourselves frequently mis-identifying the source of the problem and having to make multiple tries at correction. This is the same theme as we encountered earlier with the Quine-Duhem thesis—when combined information results in a problem or disconfirmation, we're not sure where to "aim the *modus tollens*". This would predict that neural networks, if operating in this fashion, would require a great deal of experience and training, making continuous adjustments and then trying the new configuration again, before repeating the whole process. But as anyone who has watched a child learn to walk, watched a college student study flash cards for an exam, or has tried to learn to hit a tennis ball or play a guitar knows: our brains often require a *great* deal of trial and error to learn something.[99]

The similarities between the detective operating at an agent level, the evaluation of beliefs at the doxastic level, and the activation and structure of networks at the neural level seem too strong to ignore. Indeed, the structures and properties of neurons and neural networks that neuroscience and artificial neural modeling are discovering seem quite well suited to exactly the

[99] There are also, of course, cases of more rapid learning. For example, learning new facts might, in some circumstances, require only one exposure. However, these "one-shot" cases are likely made possible by the sophisticated processing of other background components and networks, ones that themselves likely required extensive training and reconfiguration to develop and optimize.

kinds of calculation and assessment that the justificatory approach I am suggesting would predict we should observe at the biological level.

One example of the success and power of this low-level approach is found in the work of Paul Thagard (1992, ch.4). Thagard has successfully programmed an artificial connectionist network ("ECHO2") that evaluates competing scientific theories, based on coherence, simplicity, explanation, and so forth. He has tested it using scientific theories like Darwinian evolution and the meteorite impact causing the dinosaur extinction, and finds that his model yields the seemingly correct epistemic judgments about these theories versus their competitors. The power of this connectionist model in doing what scientists do provides support for the empirical feasibility of my approach, since if a computer system making use of parallel constraint satisfaction and the other techniques associated with connectionist models of neurological systems can be made to successfully perform these operations, it stands to reason that our brains, made up of the same type of connections, may well be engaging in similar operations.

Overall then, after incorporating the subdoxastic elements of our cognitive system, we arrive at the following definition:

> **(Subdoxastic/Neural) Endo-Reliability:** The tendency of a neural network to output processed information that stands in the right relationships to the outputs and processing of other neural networks is the endo-reliability of that particular network.

This yields the following formulation as our new theory of justification:

> **Endo-Reliabilism:** *S* is justified in believing that *P* at time *T* to the extent that the process of belief formation from which *P* results is neural-endo-reliable.

4.5 What About the Other Theoretical Virtues?

Before responding to objections in the next chapter, I want to discuss another interesting avenue of potential expansion of the theory developed so far. This will be easiest to consider if we return to thinking about the relationships between beliefs, propositions, scientific theories, etc., and leave the neuroscience aside for the time being.

The critical role that coherence relationships play in the above theory will undoubtedly worry some readers, likely because of a widespread belief that coherence-based theories do not have adequate resources to perform all the work that is expected of them. Things are not as bad as they seem, however. Because the theory of justification described here incorporates relations between conscious beliefs and subdoxastic mental states, and because of its holistic nature, I believe that the endo-reliabilist can also help themselves to the considerable epistemic powers of the other theoretical virtues commonly utilized in scientific reasoning. Coherence is usually itself classified as one of the theoretical virtues, since when deciding between two theories, it is thought that, *ceteris paribus*, we should prefer the theory that is more internally coherent.

I suspect that, when we talk about the various theoretical virtues, such as explanatory power, empirical adequacy, predictive power, fecundity, conservation, simplicity, etc., we are actually picking out very specific coherence relationships. While a full treatment of the reducibility of all the myriad theoretical virtues will have to wait for another time, I sketch some of these ideas here. The general idea is that, when assessing any of these theoretical virtues, our cognitive system is always comparing the theory under consideration with *other sets of information*, described below, and that at their core, it is going to be the coherence between the theory and the other particular information sets that produce the theoretical virtue. Let's examine several examples of these virtues at work for a theory, T_1.

Explanatory power- The explanatory power of a theory is often described as a result of the number (and perhaps kinds) of observations or instances that can be successfully subsumed under the theory. So, the coherence relationship at work here will be the degree to which the various observations and instances previously encountered (and currently remembered!) cohere with T_1, such that there are many observations that each have very high probability conditional on T_1.

Predictive power- The degree of predictive power will be a result of the number of cases, situations, etc. that cohere with the information in T_1.

Empirical adequacy- This is one of the primary constraints on theory selection. However, since, as I argued previously, we cannot observe empirical reality outside of our subjective experience, empirical adequacy must be a function of the fit between T_1 and all of the beliefs and information about empirical reality to which we are currently committed.

Fecundity- The fecundity of a theory is a measure of the how fruitful the theory has been at leading to breakthroughs and innovations so far, as well as the anticipated fruitfulness of further research which could be conducted if T_1 is accepted. In order to anticipate the further research, we must be able to envision (at least some of) the avenues for future research, and we are considering how well those avenues cohere with T_1.

Conservatism- The virtue of conservatism is a measure of how much of our previous theoretical framework can be retained if we accept T_1, or, in other words, how much of our previous theoretical commitments is coherent with our new theory.

Simplicity- Simplicity, I'm afraid, is a different matter. There seems at first to be no reason to think that "either reality is simple or it isn't" (although an idea of "maximal simplicity" is an interesting one to consider!), since we do not seem to have a reason to think that reality could not be complex indeed, and thus require a non-simple set of beliefs to accurately capture its nature. Instead, the reason for simplicity as a guide to truth (or pragmatic success) is an instrumental one—if we know that coherence (taken to the extreme, which, of course, we could not *actually* accomplish) is a guide to success, then we should be very concerned with *testing* for coherence, and it is here that simplicity definitely becomes a virtue. Given that we have limited epistemic and cognitive resources to allocate, we *should* always prefer the simplest of the otherwise equal, empirically adequate alternatives, since it will make coherence testing easier. Additionally, scientific practice has a long history of preferring the simpler theory, and this has contributed considerably to its success.

Hopefully, this provides a revealing and interesting glance into a potential extension and application of endo-reliabilism. As I've stated above, I suspect that the other theoretical virtues all pick out particular coherence relations, and so are ultimately and completely reducible to

coherence. With the addition of these theoretical virtues as resources for endo-reliabilism, I think the true power and flexibility of the theory becomes clear.

Admittedly, many purposes are better served by treating the theoretical virtues as distinct from one another, and this is how standard scientific use of the concepts proceeds. Scientific practice uses the theoretical virtues to justify its theories, and its selection of one or the other in cases of competition.[100] If my earlier, Quinean-style claims are correct, and our beliefs (and perhaps other cognitive states) are best treated as scientific theories (in a broad sense), and we take seriously the call from methodological naturalism to utilize our best scientific practices in our epistemology, then it is a positive result indeed that our individual beliefs are also justified on the basis of their coherence and embodiment of the other theoretical virtues.

[100] In the Appendix, I develop a related model that I think offers a better and more precise way to understand and utilize the theoretical virtues in theory evaluation. This approach is not dependant on endo-reliabilism, and may have numerous applications for scientific practice independent of my other arguments and views.

Chapter 5: Defending Endo-Reliabilism

At this point, it should be clear how endo-reliabilism meets the *desiderata* established in chapters 2 and 3. To review briefly:

1. Endo-reliabilism is naturalistic,
2. Endo-reliabilism focuses on process and causal history,
3. Endo-reliabilism is internalist, as I have defined it, and only utilizes relata internal to the cognitive system (in order to avoid many of the objections to externalism),
4. Endo-reliabilism is informative about, and makes use of, the subconscious/subdoxastic states that constitute a sizable and important share of human cognitive processing,
5. Endo-reliabilism is pragmatically oriented and constrained,
6. And, endo-reliabilism has normative force (telling us at least how we *ought* to form and modify beliefs given our goals.)

Along the way, we have also seen the considerable support for this theory offered by (recent) experimental findings and approaches in the sciences. The purpose of this chapter is to consider several objections to the theory that either have appeared in the relevant literature, have been raised by previous readers, or that I anticipate are likely to be raised, and discuss how an endo-reliabilist might respond to them.

5.1 Objections Based on Intuitions

Many readers may find that the theory I have described clashes with their intuitions in some way or other, and this will often be cited as evidence against the theory. Whether these conflicting intuitions actually provide a challenge to my theory or not (and if they do, to what extent), is, however, an open question. The role of intuitions in philosophical theorizing is an important meta-philosophical topic, and there is a considerable amount of exciting work being done on the topic at present.[101] One recent trend is that quite a few naturalist philosophers find ourselves relegating intuitions to a far less important role than has been ascribed to them in traditional philosophical methodology. Let us look briefly at two reasons why I (and others) believe this demotion is appropriate.

First, I follow Kornblith (2002) in thinking that the actual investigatory target of epistemic research is *not* our *concepts* of epistemic terms (like knowledge and justification), but rather the *phenomena* in question themselves.[102] Kornblith points out that, when a chemist sets out to learn something (say, some property of an acid) they are not interested in the current *concept* of acid and its entailments, they are interested in the acid itself and its properties. This is why they engage in empirical work aimed at testing multiple hypotheses about the chemical in question, and do not *merely* consult their *intuitions* about it. After all, history is filled with an overwhelming number of cases where the intuitions associated with the leading conception at the time were utterly mistaken. Similarly, it may be that our concepts about things like the nature of

[101] Examples include Prinz (2008) and Williamson (2007).

[102] Given Kornblith's realist leanings, I think it might be more accurate to say that he thinks it is the *noumena* in question that are the actual targets of our investigations. My view is that we aim at empirically adequate theories and models of the "things themselves", but that what we are really investigating is some subset or other of our experiential world. This is still importantly different from the traditional view that our epistemic theorizing aims at analyzing the relevant *concepts*.

minds, the nature of the external world, knowledge, and justification are currently mistaken, and so *intuitions* about these concepts are simply not relevant in the way that typical analytic philosophical practice has presupposed.

Of course, some of this depends on what, exactly, intuitions are. Some philosophers sympathetic to the rationalist tradition have advocated that philosophical intuitions somehow give us privileged access to the truth, but there are numerous questions about how this "privileged access" would work, and seems to be exactly the type of claim that the naturalist approach eschews. Instead, a naturalist might well view intuitions as informational outputs from certain cognitive systems, distinguishable from other conscious informational outputs by the fact that the inputs and informational processing involved in the formation of the intuition are not consciously accessible to the agent. In other words, intuitions are the result of "dark processing" in the brain, seeming to the agent to just "arise out of nowhere", because the agent cannot reflect on the cognizing that led to the intuition. Because of this view, Kornblith and others suggest that intuitions are only useful when we confront a new topic or question, one for which we lack any theories or evidence. In these cases, intuitions may be useful to help steer our investigations and theory construction, but once we have generated theories that we *can* evaluate using established best practices for theory selection, we should move away from utilizing the starting intuitions as much as possible. Once this point is reached, if a theory and an intuition conflict, then so much the worse for the intuition. Given this view of intuition and its proper role in philosophy, my dissertation has aimed to build from information about our epistemic situation and about what our science tells us happens within our brains when we form or revise beliefs, without giving much weight to our (largely pre-theoretic) intuitions about how we "feel" justification should be seen.

A second important challenge to intuitions and their use comes from well-known results in experimental philosophical work, such as the classic 2001 paper "Normativity and Epistemic Intuitions" by Weinberg, Nichols, & Stich (2001). In this paper, experimental data suggest quite strongly that intuitions that were once thought to be universal actually vary according to factors such as culture, gender, socio-economic status, etc. If the intuitions that analytic philosophers have traditionally cited are not universally shared, and indeed might even be a minority view held primarily by so-called W.E.I.R.D. (Western, Educated, Industrialized, Rich, Democratic) people (Henrich and Norenzayan, 2011), then it is hard to see how they can actually play the critical evidential role that was assigned to them by traditional analytic methods. Given that intuitions (our traditional philosophical "bread and butter") are now suspect, an objection to my view that rests *solely* on intuition is not as forceful as it might at first seem.

5.2 Objections Involving Coherence

5.2.1 Simple Set of Beliefs as Most Coherent

One of the classic objections to coherence theories of justification is that high degrees of coherence may be easier to obtain with a smaller set of beliefs than a large one. Indeed, maximal coherence can be easily obtained by adopting any one belief and no others (since any belief is perfectly coherent with itself, and its conditional probability is 1.0— *P* must be true given *P*).

Even if this is true of coherence-driven systems *in theory*,[103] however, this does not apply to actual human epistemic agents because we cannot cognitively disable the stream of sensory

[103] Of course, whether or not this is true of coherence theories is debatable. For example, BonJour (1985) attempted to address this objection and show that some coherence theories can provide a satisfactory response to this objection.

input that we constantly receive. By virtue of our neurological systems' functioning, we have sensory perceptions across time, and until they are subsumed under *some* organizational structure, they yield rampant incoherence (for example, just one instantaneous visual experience can yield "redness and not-redness" in different parts of the visual field). Our contingent biological nature results in an unavoidably large set of information, and so the comparisons that are relevant to the current topic will be between large sets with varying degrees of coherence, not between large and small sets.

5.2.2 Coherence Is Not a Guide to Truth

Another related and frequent objection is that coherence is not a guide to truth, since there can certainly be highly coherent sets of false beliefs (e.g. Klein and Warfield, 1994 and 1996, and BonJour, 1985). Consider the following set of beliefs: {unicorns are horned horses, horned horses exist, unicorns exist}. This set coheres very well, but contains little in the way of truth. So we might worry that coherence-driven theories could push us to prefer false but coherent sets over (more approximately) true, but potentially less coherent sets.

First of all, I have already argued that truth should not be seen as the central goal of epistemology (see Chapter 3, section 3), so I think the better question is whether it should worry an endo-reliabilist that a set of false, highly coherent beliefs could be labeled as more highly justified than another less coherent set of beliefs that includes more true beliefs.

On my theory, the relevant coherence relations result from the entire system, all of its processes, and its goal states, so if all of this is coherent, even the low-level, subdoxastic states, then it makes sense to say the belief in question is indeed justified (even if completely false). If we imagine a computer system or robot that has been "misprogrammed", and so makes chronic mistakes, but performs perfectly within the parameters it was programmed under, then I would

think that its performance to the best of its ability, in accordance with its nature, is exactly what should earn a label like “justified”. This important distinction between having things right (true) and having them right to the best of the system’s ability is exactly why we have different concepts for truth and justification.

5.2.3 Confusion of Epistemic and Prudential Justification

Some readers may be concerned that, by rejecting the idea that epistemology fundamentally aims at arriving at true beliefs and instead advancing a pragmatic approach, I have simply collapsed the notions of epistemic justification and prudential “justification”. This, however, is not the case, as there are situations where, on my theory of epistemic justification, an agent will be justified in one of these ways, but not the other. This, of course, means that they cannot be identical if their ascription differs in the same case. Let us consider one such example. One of the standard ways of describing the difference between epistemic and prudential reasons in a classroom setting is to ask students to imagine a case where they are in a completely empty room and are then told that if they can really *believe* that there is a giraffe in the room, they will be paid an exorbitantly large sum of money. On the (in my experience very safe) assumption that students want more money, the potential payment involved would provide the students with a compelling *prudential* reason to form the belief that a giraffe is present. However, the conflicting sensory data that decisively shows that no giraffe is present, and perhaps other factors like the geometrical reasoning that makes clear that no giraffe could fit inside a room of the size in question, show that even in the face of an overwhelmingly strong practical reason to believe the false proposition in question, the students would not be epistemically justified in actually believing it.

The theory of justification that I am advocating yields the same, intuitive, result in this scenario. While the monetary reward provides a strong prudential reason for forming the belief, the theory's focus on the *process* that formed the belief in question shows why it is not epistemically justified. A process which, just to give one possible description, *forms a specified belief in exchange for the promise of money* will not be endo-reliable, since the beliefs it yields will tend to cohere quite poorly with the rest of the belief framework. Even though it may cohere quite well with the agent's conscious beliefs (such as "I ought to do that which will give me additional money and believing *P* will result in me acquiring more money"), the total belief framework also incorporates components such as the subdoxastic perceptual states that result from the sensory experience of the empty room, etc, and the lack of coherence at this holistic level of the beliefs that result from the belief-forming process in question means that the particular belief in question is unjustified.

Similarly, there will be cases where the endo-reliability of a process causes a belief to be justified, even though holding that belief will make the agent sad or prudentially worse off in some other way. These will be cases where the agent is epistemically justified but prudentially *unjustified* in holding the belief in question, and thus further demonstrate that the theory here on offer does not collapse epistemic justification into "merely" prudential justification. We still need both traditional categories, I just think, as I have argued, that we have been wrong about what *constitutes* epistemic justification.

5.2.4 Why Not Just Be a Foundationalist?

Why prefer a coherence-driven account of endo-reliabilism and not be a foundationalist? First of all, I agree with Pollock and Cruz (1999, p.86) that we should reject what they call "the doxastic assumption" frequently made by foundationalist theories, which holds that the

justification of a belief is a function of *only* the other beliefs that one holds. This assumption is faulty because it fails to take into account the critical role that subdoxastic states play in our cognitive economy. On the other hand, it may be possible to have a foundationalist theory that does somehow incorporate subdoxastic states, and this would be a step in the right direction. However, the second problem I see for foundationalism is inherent in the very structure (so to speak!) of the theory. Foundationalism asserts that some set of our beliefs (usually low-level sensory beliefs) are privileged (basic, given, etc.), and so are capable of serving as the foundation for all the rest of our beliefs. I reject this idea, and instead think that Quine painted the better picture in his portrayal of our system of beliefs as a web where nothing is foundational, and any piece of the web *could* in principle be changed and revised. Foundationalists tend to simply assert that the foundations are brute or primitive, and so do not themselves need to be justified (or are somehow "self-justifying"), but this seems to end up rather *ad hoc*, especially if, as naturalists, we aim for a theory that bottoms out in *completely* non-evaluative terms like Goldman (1979) suggested.. There are also empirical reasons to be suspicious of the foundationalist picture, since we are now discovering that recurrent neural network-style processing is extremely prevalent in the brain, including lots of feedback from downstream processing to earlier modules, and this shows that even "foundations" like how we *see* what is in front of us can be changed. Part of what makes a professional baseball player more effective at hitting a fastball is that, because of all the extra synaptic training the player's cognitive system has received, the different categories of kinds of pitches the player has learned, and the extra pressure of having one's career depend on the outcome of swings of the bat, the player *literally* sees the ball differently. If even the low-level cognitive ability to detect edges and motion is affected by what happens elsewhere in the cognitive system, then it seems that nothing is safe

from these potential effects, and so nothing can adequately serve as the privileged class needed to provide a proper foundation.

5.2.5 Why Not Just Be a Coherentist?

Given all the work being done by the coherence relations in my theory, why bother retaining the focus on process and causal-history, instead of developing a modified or extended "BonJourean-style" theory? I do think coherentism with subdoxastic *states* included might make for a powerful and capable theory—one that may have the resources to avoid or handle many of the classic problems with coherentism (although this will likely exacerbate the "computational load problem" that coherentism already faces, discussed below). However, I remain committed to the view that the history of how a belief or other cognitive state came about is extremely important to the justificatory status of that belief. It seems that it is possible that a defective, non-endo-reliable process like wishful thinking could generate a belief, which then *happens* to stand in just the right synchronic relationships with other beliefs (or subdoxastic states), and could end up identified as justified by a theory that does not include the proper diachronic, causal-historical focus. This is clearly a problematic case of "epistemic luck", and I think shows that a belief can be considered justified if and only if it results from the right kind of *process* with the right features.

Naturalists have another reason to insist on this focus on cognitive processes. Our best science has found it extremely fruitful to investigate cognitive processes and the neural networks that comprise them, and so, to the extent that we take ontological naturalism seriously, philosophical theorizing should seek to construct theories that employ them as central components.

5.3 Objections Concerning Naturalism

5.3.1 Incompatibility of an Internalist Framework with Naturalism

Some readers may be wondering about my invocation of skeptical arguments to motivate epistemic internalism, while simultaneously appealing to naturalism and scientific success. After all, how can a pragmatist, who is uncomfortable with externalist theories and their connection to truth, talk about things like subdoxastic states, empirical findings about the neural workings of the brain, and the amazing past successes of scientific practice? The answer, quite simply, is that one need not be a scientific realist to help oneself to these fruits of scientific labor. Whether or not science discovers things that are actually *true* (or even approximately true), it does seem to have been remarkably successful at generating the most empirically adequate theories and models available, and these in turn seem to offer the best explanations for the various phenomena we encounter in our experiences. The naturalistic world-view is remarkably internally coherent, and, as evidenced by the many medicines, comforts, and joys afforded us by our technological developments, seems to provide us with a toolset incredibly well suited to helping us attain our goals and fulfill our desires. A pragmatist, even if nothing more than a brain in a vat, would do quite well to take science's methods and findings seriously, as they seem to provide the most useful and effective tools and data points upon which to base one's web of beliefs.

5.3.2 Epistemic Probability—Is Endo-Reliabilism Really Naturalist?

In "What is Justified Belief?" (1979/1994, p.106), Goldman lists terms such as "is probable (in an epistemic or inductive sense)" as evaluative terms that ought not to appear in a properly naturalist analysis of justification. One might worry that my theory, drawing heavily on the probabilistic relationships outlined in Chapter 4, makes use of exactly this kind of probability, and so has not met the demands of naturalist approach. However, Goldman does

allow use of probability "(either in the frequency sense or the propensity sense)" (Goldman, 1979/1994, p. 106). The probabilistic terms used in my theory are objective probabilities located at the level of neural events that are certainly allowed in a naturalist's science-driven discourse.

5.3.3 "I'm Still Not Sold on Naturalism"

While we have seen (in Chapter 1) some of the reasons that many philosophers have been adopting naturalistic approaches to epistemology, I wish to look quickly at some possible reasons why someone might resist a naturalistic approach.

First, I suppose it is possible that one might disagree that science is successful. This strikes me as an extremely untenable position in the face of the evidence. Science has allowed patients born without vision to be able to see later in life, it has allowed us to clone sheep, predict tsunamis, send probes beyond the edge of our solar system, and construct a world wide web of computers that stores tremendous amounts of information and has such an astonishing number of interconnections that it rivals the human brain! I must confess that I do not have an argument against an individual that wholly rejects science; in fact, given the radical difference in intuitions and worldview, I do not think any meaningful dialogue with such an imagined person would be possible.

Second, one might accept that science is successful, but that is does not (necessarily) get at the truth. This might well constitute an objection to more robust forms of naturalism, such as Kornblith's view that knowledge (and perhaps other targets of epistemic inquiry) are natural kinds and so should be investigated by science in order to discover the truth about these kinds. However, acknowledging that science is successful is adequate for launching the naturalist's project. Indeed, I argued in chapter 3 that our epistemic theorizing should be "pragmatically

constrained", focusing on the success of our efforts rather than on the objective *truth* of our discoveries and beliefs.

Third, someone could well think that science is successful when applied to its appropriate topics, but that (at least some of) the epistemic concepts we are interested in as philosophers are just not part of the domain for which scientific investigation is appropriately employed. This is a more difficult objection to respond to, and seems to be by far the most common reason that some philosophers resist naturalistic approaches. However, I think there are at least four responses that can be given here. The first response is simply to claim that it is logically impossible for anything in the actual world to be outside of the realm of scientific investigation because science investigates everything that is actual, and so tautologically, anything in the world is properly understood as an appropriate target of scientific investigation. The second possible response would be to acknowledge that it is *conceptually* possible that some things are outside of the realm of scientific investigation, but that this is metaphysically (or at least nomologically) impossible. The third response is to acknowledge the intuition that the objector has, but then provide an error theory for the intuition. For example, it makes sense that early in a scientific investigation, the success of the investigative projects looks unlikely or even impossible because of the limited information and lack of a plausible explanatory theory. But as the investigation proceeds and more information and possible explanations are accumulated, the success of the investigation comes to be seen as possible, probable, or even inevitable. There seems to be a great deal of evidence for this throughout history. In the Middle Ages, suggesting allocating resources to investigate scientifically what is causing the strange behaviors of those people

"possessed by demons" would be viewed as silly and wasteful, not to mention blasphemous.[104] It appeared at the time that the proper explanation for the state of the mentally ill was not included in the categories for which scientific and technological investigation was appropriate. We now know better, since psychological and psychiatric care has made tremendous progress in explaining and treating these same psychological ailments. Given how often this pattern has repeated throughout humanity's history, we ought to be by now very wary about pronouncing anything beyond the "realm of science" *simpliciter*, and should at most call something beyond the *current* reach of science. The fourth response is the weakest but also I think the most persuasive, which is to point out that the alternative, non-scientific, approaches to investigating these philosophically interesting topics have had centuries (or even millennia) to work on them, and have not made the progress for which we had hoped. As a result, instead of continuing the same approaches and hoping for different results, we should take a break from those approaches and investigate the newly available alternatives. Every day, modern science generates tremendous amounts of new data and new theories— and this constitutes information that no human has previously had available. It would seem a better use of our time and energy to attempt to employ the new resources to the old problems than to keep trying the same approaches as Plato, Aristotle, Descartes, Kant, and so forth, and hoping that *we* can do what *they* could not.

The fourth reason someone might have for resisting naturalism is really more of a complaint, which is to acknowledge that there is a case, *prima facie*, for pursuing the scientific investigation of epistemic concepts, but despair about whether this route will furnish us with a satisfactory answer. Most often, this is encountered in people that agree that science is the appropriate tool for investigating the world, but worry that science cannot provide us with

[104]On December, 5th, 1484, Pope Innocent VIII issued a decree that witches actually exist, and made questioning this assertion heresy (Haught, 1990).

theories that are robustly normative like we expect and hope to find in ethics or epistemology (e.g. Kim, 1988). Here, I see two possible responses: first to repeat the response above to the third objection and point out that we do not yet know what this route can accomplish (and so perhaps robust normativity is attainable), or second to remind the objector that lots of things do not turn out as we hope, and that not liking an outcome neither makes it false nor means that it is not the best alternative from among a set of non-ideal possibilities. The normativity concern at work here is discussed further below.

5.4 Cherniak and Computational Load

Christopher Cherniak (1984, 1986) observed that many epistemic theories, including most versions of coherentism, place demands on rational epistemic agents that cannot possibly be met by actual humans. Specifically, the process of checking the consistency of a set of beliefs is an enormous computational task for a surprisingly small set of beliefs. While human brains are quite powerful in their computational abilities, Cherniak shows that "[a] surprisingly small and basic fragment of the ideal agent's deducing ability seems by itself to require, for a finite set of simple cases, resources greater than those available to an ideal computer constructed from the entire universe" (Cherniak, 1984, p. 245). One of the key theoretical components identified by Cherniak as necessary in order to avoid this seemingly decisive objection is the utilization of heuristics, which allow a lightening of the computational load, albeit at some cost of accuracy and thoroughness. By focusing on processes and procedures of belief formation instead of the individual beliefs themselves, Endo-reliabilism essentially provides a compression of the processing required for evaluating beliefs. Endo-reliabilism's notion of reliability-defined-endogenously means that we are justified in believing something that results from a reliable process (defined endogenously) without having to calculate coherence with every other belief.

And, rather than just pointing to "some heuristic", endo-reliabilism's close relationship with psychological and neuroscientific theories means that we may eventually be able to see the probabilistic formal machinery (like that suggested by Bovens and Hartmann) realized at the neural level, allowing us to see how the brain *is* the heuristic.

Another promising route for responding to this worry may be provided by dynamic systems theory, which has shown that in a complex dynamical system there needn't be independent assessment steps in order to have adjustment to the system. Instead, in some systems (such as the oft-discussed "Watt governor"), the assessment and adjustment function is built into the system itself. Perhaps the brain is a similar system and does not require any higher-order modules that assess the coherence and endo-reliability of the brain processes because the assessments are self-adjusting via synaptic strengthening or weakening. Further research on this idea will involve studying the mechanisms of Hebbian learning as a potential method by which the brain's computational demands can be compressed, and thus be met by actual, non-ideal human brains.

5.5 Mental Content

One might ask about how beliefs get their content, and about which theory of content endo-reliabilism employs. While it is not customary for theories of epistemic justification to specify a particular theory of mental content, I recognize that there are numerous ways in which the issue could be relevant to my project. For example, if an externalist (broad) version of content (perhaps a causal theory such as to Dretske's 1981 Causal-Information Semantics, Millikan's 1984 Teleosemantics, Fodor's 1987 Asymmetric Causal Dependency Theory, or Rupert's 1999 Best Test Theory) is correct, this could considerably weaken the force of some of my arguments in favor of skepticism in Chapter 2 and for pragmatism in Chapter 3, and could

potentially have consequences for the types of relationships that will hold between beliefs and other cognitive states. My theory is more easily seen to be compatible with an internalist (or "narrow") theory of content, since these theories typically do not depend on the environment within which an epistemic agent is situated. Examples of these approaches include Conceptual Role Semantics (Block, 1986) and Hyper-Dimensional State Space Semantics (Churchland, 2007, 2012). Of these, I find Churchland's theory particularly attractive, both on independent grounds and because of how well I think it complements endo-reliabilism. The attempt to provide meaning and content while eschewing traditional propositional structure has particular promise for elucidating the states and coherence relationships my theory employs at the neural level.

Still, keeping with the spirit of skepticism for which I have advocated, I wish to remain agnostic on this admittedly important philosophical topic. Given that one of the main functions that content plays is to specify under what conditions the belief or other mental state in question is true or false, and that my theory downplays the relevance of truth and falsity, in this way at least it is not extremely important for a theory of content to be specified. But there are other ways in which it is likely to matter. My hope is that either an internalist theory turns out to be correct, or that suitable modifications could be made to my arguments and the theory developed here to accommodate an externalist notion of content. I think this is likely, especially given that the coherence relations that are of central importance to my approach could just as easily hold between beliefs with external content, but exploration of these issues will have to wait for another time.

5.6 Normativity

One of the specific difficulties many readers are likely to worry about is that if the subdoxastic processes are as important as my theory suggests, it seems difficult to reconcile my theory of justification and the normative role that we often think such a theory ought to play. It is typically assumed that, like in ethical theorizing, a correct theory of justification must offer guidance on what we *ought to do*, and not just describe what we in fact do. Of course, the objection that naturalist theories of justification fail in this regard, and are not sufficiently normative, is nothing new to naturalists.[105] However, many people might worry that, by divorcing my naturalist theory of justification from truth, I have made the situation even worse, and that there is no way for the theory here offered to incorporate any normativity.[106] While I think that an answer to this has already been delivered in the section on pragmatism (3.3), the topic is worth discussing further.

Many defenders of naturalist theories face criticism that their theories fail to be genuinely normative, and are unable to give a robust enough account of what we *ought* to do. This sentiment is especially strong among advocates of a deontological notion of justification. According to this approach, which is often paired with access internalism (as traditionally defined), it is essential that justification be robustly normative and determine what we should believe, what is permissible to believe and what is not, and what is required to satisfy our epistemic duties, obligations, and responsibilities. There are, however, a number of reasons for rejecting this conception of justification, the strongest of which, to my mind, is that it depends on the traditional assumption that we are epistemic agents that can choose what to believe, and

[105] See Kim (1988) for an influential example of this complaint.

[106] For example, when discussing normativity, Kornblith states, "that any account of epistemic evaluation which does not give truth a central role to play is inadequate" (1993, p. 372).

therefore regulate our epistemic actions. By this point, however, we have seen numerous reasons to reject this picture. I have argued that the interesting and important epistemic activity happens at a level far below that of an agent (as traditionally conceived), and that many or all of the relevant J-factors are too far below the level of our conscious access (if there even is such a thing!) to be monitored and regulated in the ways required for a deontological approach. I will shortly argue that my theory can, perhaps surprisingly, deliver *some* of the regulative control that advocates of deontological justification hope for, but I think it has been shown that even a total failure to do so would not be as problematic as it might at first seem. Even the complete rejection of deontological justification and any possibility of regulative control would likely not cause concern among the primary audience at whom my theory is aimed: epistemologists that, because of their naturalistic commitments, have embraced an externalist version of process reliabilism.

There is already an "intuition shoving match" that often happens around the topic of normative requirements, since many naturalists seem to find it clear, obvious, and intuitive that the normativity that naturalism does afford is adequate, while non-naturalists frequently find it just as clear, obvious, and intuitive that it is not. My primary aim here is to show that my theory *can* deliver adequate normativity to satisfy the naturalist.

Let us then proceed by examining what normativity is commonly found in naturalistic theories, and from where it originates. Kornblith, a paradigmatic naturalist by almost any measure, asserts that "epistemic norms are a variety of hypothetical imperative" (1993, p. 359), such that *if* we have *desires*, then we should acquire our beliefs in a particular way. He argues that

> epistemic evaluation finds its natural ground in our desires in a way which makes truth something we should care about whatever else we may value. This provides us with a pragmatic account of the source of epistemic normativity, but an account which is universal and also allows truth to play a central role. Pragmatists have typically suggested that epistemic evaluation will have little to do

> with truth; but if I am right, it is for pragmatic reasons that truth takes on the importance it does in epistemic evaluation. (Kornblith, 1993, p. 373)

There are two important things to note here. First, it is having *desires* (roughly what I have been calling goals and goal states) that constitute the antecedent of the hypothetical imperative and generate the normative force that attends good epistemic practice. On this point, I am in complete agreement, and think that Kornblith has made the correct diagnosis. Second, the value of truth is pragmatic (or instrumental), not intrinsic.[107] Here again, I agree with Kornblith, but I instead side with the pragmatist conclusion that he rejects. I agree that, *to the extent* that truth has value, it is pragmatic or instrumental value, but as I argued in Section 3.3, I think we have consistently overestimated the value of truth, and should instead focus on pure pragmatic value by orienting our epistemic theorizing to it. If normativity does not derive from the truth-orientation found in other epistemic theories, then we can completely jettison it if another superior option is available. I have argued that endo-reliabilism, and the pragmatically constrained and motivated coherence relationships it employs, do offer a *better* and more theoretically virtuous account of how we should form beliefs in order to best attain our goals.

Endo-reliabilism gives us a way to evaluate a cognitive system's beliefs by examining the endo-reliability of the cognitive processes that formed and maintain them, and this, in turn, yields a way to identify which beliefs the system *ought* to hold (namely, beliefs which result from processes with high degrees of endo-reliability), and which it *ought not*.[108] Because all of

[107] Maffie (1990, p.333) expresses a similar view, stating that "epistemology is normative only within the framework of instrumental reason and…its normativity is parasitic upon that of the latter."

[108] Of course, I am not suggesting that there is one set of beliefs that *everyone* should hold. I take the arguments in section 3.3 to have shown that *which* beliefs a cognitive system *should* hold will be influenced by that particular system's goals.

the relata are internal to the cognitive system, it can engage in this evaluation internally, allowing endo-reliabilism to be employed by cognitive systems to cognitively regulate their belief adoption and modification practices (which we will examine below). The theory entails that we have cognitive access to the evidence and support for *why* our processes are reliable, in a way that does not require allusion to states of affairs outside of our epistemic situation, nor restrict us to things that are consciously accessible. The J-factors that determine a belief's justificatory status are all internal to the system, being completely comprised of mental states, cognitive processes, and the inter-relations between them described previously, but they still provide a normative epistemic standard that *ought to be followed* if we want to attain our goals. And it is, I believe, an inescapable part of our nature that we do wish to attain our goals (indeed, it is built into what a *goal* is that one is driven to attain it). For whatever reason, perhaps through nothing more than a grand cosmic accident, our systems are such that they seek continued existence, avoiding painful experiences, seeking pleasurable ones, and so on. Perhaps these goals and desires result from evolutionary selection pressures, since any organisms like us that lacked these goals would have died out long ago. Or perhaps an omnipotent evil demon thought it would be the most fun to create and deceive a being with these particular goals and desires included. Either way, these goals and the other more specific ones we set for ourselves are *ours* and it is from them that epistemic normativity derives. Whether we are brains in vats or inhabitants of an external world that is exactly how it appears in experience, we still have every reason to pursue the particular goals that we have, and as effectively as possible.

Concerning effectiveness, we saw earlier that one of the purported advantages of an internalist approach to justification is that such an approach allows an agent to *regulate* their beliefs. This capability, pragmatically speaking, ought to increase our effectiveness at attaining

our goals. So how does this work on endo-reliabilism? Regulation will be operative in two different ways, at correspondingly different levels. First, at the subconscious level, I think that regulation using the endo-reliability metric is widespread in an automatic, systemic way. I have argued that it is a feature of connectionist-type neural networks that a part of the network that is not meeting the processing needs of the system (by being endo-reliable), will be "devalued" by the system over time, by having the synaptic weights associated with its outputs decreased. For an illustration of this idea, imagine a case where a scratch on a subject's cornea suddenly produces a corresponding blind spot in the visual field. As time goes by, the endo-reliability of the associated proximal process will decrease because its output will remain constant while the other processes assessing the information from nearby, non-blind parts of the eye will be changing in myriad ways. This means that its outputs will be far less coherent with the outputs of the other processes, and over time, the system will adjust to this "endo-unreliability" of the affected process. The brain has shown itself to be incredibly flexible in handling damage, in some cases by completely bypassing and replacing the processing of large areas of brain, such as in the case of severing the corpus callosum that links the two hemispheres of the brain—a medical procedure typically performed to stop the spread of seizures (Bermúdez, 2010). If the brain automatically adjusts in these cases, with corresponding adjustments made for well-functioning, highly endo-reliable processes, I think this provides an adequate account of how our cognitive system regulates itself in many of the ways that epistemology has thought that it does.

While this neural-synaptic regulation is not under our direct control (I will argue momentarily, however, that we can exert considerable *indirect* control), it helps provide an explanation of the differences between when our cognitive systems do things well and when they fail to do so, as well as for why some people's systems are better at some things than others. For

example, there are clearly epistemic differences between the beliefs held by a delusional, paranoid schizophrenic and held by a more successful (perhaps "virtuous") reasoner.[109] As I have argued, the important epistemic differences between these two cases are found in the differences in the probabilistic coherence relationships of their cognitive processing, *not* in the number of truths attained. A hallucination, while traditionally characterized as epistemically defective because it is false, is also highly incoherent with the outputs of the other cognitive processing, and so is defective on my view *because* it fails to result from an endo-reliable process. A successful reasoner, on the other hand, is successful or epistemically virtuous precisely because she has and employs endo-reliable processes. This also highlights what makes experts and expertise valuable. By becoming experts or developing expertise, the relevant cognitive systems have added more information to the cognitive web, identified more coherence relations between the information, and have practiced and trained the cognitive processes at consistently generating outputs that are highly systematized, coherent and stand in the right probabilistic relationships. It is this increased and improved training, occurring at the subconscious neural level through synaptic strengthening or weakening, that allows an expert botanist to easily and automatically categorize plants, allows a skilled logicians to "just feel" that there is a defect in an argument and then identify it, allows a studied historian to recall more

[109] Although, perhaps in an extreme enough situation, a "delusional" world-view *could* be justified. Imagine a case where someone is in a horrifically bad situation, perhaps in a Nazi concentration camp or something similar, where almost all of their goals are being frustrated and changing this is completely beyond the individual's ability. In this case, the only way to pursue the agent's remaining goal of happiness might be to shift into a "delusional" experiential world where things are better. Perhaps, as a coping mechanism, what would otherwise be called insanity could be exactly the best approach and what one *ought* to do in those circumstances. Endo-reliabilism will, even here, point to better and worse ways for the cognitive system to do this.

dates and the significance of the historical events that occurred at that time, and allows a professional baseball player to "see" the ball more accurately and swing at it more effectively.

This neural-synaptic training that makes us more or less effective at a subconscious, automatic level, points the way to the second level at which endo-reliabilism allows for belief regulation. This occurs at the more traditionally discussed "agent" level, and it is here that the "self" that we take ourselves to be can exert control (albeit often only indirectly). To be clear, I am not backing away from the views I argued for earlier, nor is there anything "spooky" or "magical" happening here by thinking of ourselves as agents. Since I think we are *nothing but* our cognitive systems, when "we" exert control over the functioning of our cognitive system, it is just *parts* of our cognitive system affecting *other parts* of the same cognitive system. Still, it is often useful to talk about the particular systems that are associated with the "us" that is "conscious" and "under our control".

We saw above that the key to improving our epistemic activities, goal attainment, and level of justification for our beliefs is to increase and improve the training of our neural networks. But how can we affect this neural level? The answer is that we can put ourselves in situations that have the best chances of training up our neural nets in the correct ways. We can seek out and place our cognitive systems in environments that facilitate developing better cognitive processes, by exposing ourselves to the kind of feedback and stimulation that will force the valuable synaptic improvement and readjustment described above. We can choose environments and inputs that will "tune" our synaptic weights in the right ways, and we do this by going to classes; studying journal articles carefully; practicing skills over and over; seeking input from experienced teachers, coaches, or interlocutors; seeking out multi-modal sensory reinforcement; practicing the use of mnemonic devices and strategies; meditating to develop

increased focus; and so on. These are exactly the “tried and true” techniques for improving ourselves as thinkers and actors, but endo-reliabilism provides a deeper, and I think more compelling, story about why these things work, what is happening in the cognitive system, why it is valuable, and therefore, why we *ought to do them*.

6 Conclusions and Consequences

One of the important points made by Bishop and Trout (2005); Goldman (1986), the exploration of cognitive heuristics by Gigerenzer (1999), Gigerenzer, Todd, et al. (2012), and Kahneman (2011); and others is that our understanding of epistemic norms, rationality, and other epistemic goods needs to be informed and properly constrained by the limitations of our cognitive systems. In this dissertation, I have argued for additional limitations, particularly that our theories should take skepticism seriously, remain limited and constrained by pragmatic considerations, and only utilize J-factors that are internally accessible by the cognitive system. Taking these limitations into account, I believe that endo-reliabilism delivers the best theory of justification with the maximum amount of normative force possible.

To review, the theory of endo-reliabilism holds that *S* is justified in believing that *P* at time *T* to the extent that the process of belief formation from which *P* results is neural-endo-reliable, and the neural-endo-reliability of a particular network is its tendency to output processed information that stands in the right relationships to the outputs and processing of other neural networks in the cognitive system.

So where does this leave us? At this point, the dissertation has hopefully at least shown that endo-reliabilism is a view worthy of serious consideration and further examination within the epistemological possibility space, but let us briefly review the key ideas that led us here by examining and evaluating how the theory I have developed satisfies the *desiderata* laid out earlier.

6.1 Meeting the Desiderata

Desideratum 1: The theory should be naturalistic.

As was discussed at the beginning of the dissertation, this project has aimed to conform to the commitments, both ontological and methodological, of contemporary naturalist epistemology. Indeed, it was argued that one of the primary strengths and attractions of standard externalist versions of process reliabilism is their conformity with the naturalist framework, and that this was a key constraint on the new theory to be described. Obviously, endo-reliabilism does retain this naturalistic orientation, since it holds that it is the completely physical, scientifically investigable, processes located in the brain that confer justification. The entities and relations included in endo-reliabilism are of the same types as are currently being investigated by the relevant scientific disciplines, and so are in line with a naturalist ontology.

Endo-reliabilism also conforms to the methodological component of naturalism. It draws from and builds upon findings from several interconnected scientific fields, and even helps to identify some specific areas for future experimentation and research.

Desideratum 2: The theory should focus on process and causal history.

The emphasis on process is central to the theory of endo-reliabilism. As was mentioned, this focus is important to the successful differentiation between when a belief is justified and when it is justifiable. Endo-reliabilism holds that a belief's justificatory status is tied to the process that formed (or maintains) it, and how well that process meets the relevant conditions.

The key change that allowed us to retain this focus on process and causal history is the employment of a different notion of reliability. By utilizing an endogenous definition of

reliability instead of the typical, exogenous one found in standard externalist versions of process reliabilism, we generate a theory of justification better able to satisfy our other requirements and *desiderata*. As was discussed in Chapter 1, this focus also allows endo-reliabilism to differentiate between beliefs that are *justified* and those that are *justifiable* (which is itself arguably a *desideratum* for any adequate theory of justification). Endo-reliabilism, by focusing on the endo-reliability of the process used to form or sustain a belief, has the resources to separate beliefs that are justified, from those that result from a process that is not endo-reliable (and so are unjustified), but that *would* have been justified if a different cognitive process had been used instead.

> *Desideratum 3: The theory should be internalist, and only utilize relata internal to the cognitive system (in order to avoid many of the objections to externalism),*

As was discussed previously, many Internalists will be inclined to label the theory that I have developed as externalist because of the lack of conscious access to many, or most, of the justificatory relata. But Externalists are likely to insist that my theory is also not an instance of an externalist theory. I have argued that there are good reasons, independent of satisfying this *desideratum*, to adopt a different definition of internalism about justification, one that instead draws the boundary at the perimeter of the cognitive system. While this is a change from traditional epistemic definitions, it is endorsed by other naturalist philosophers (for example, Pollock and Cruz, 1999), and is compatible with, and perhaps encouraged by, recent work within cognitive science, psychology, neuroscience, etc. Under this new and improved definition of internalism, my theory does certainly count as an internalist theory of justification.

Indeed, on one interpretation, my theory is *more* internalist that most theories—even those typically labeled as internalist in nature! One reason that a traditional internalist might

object to traditional externalist process reliabilism, is because, by defining justification in terms of the track record of processes to produce beliefs that are objectively true, externalism places an important relatum outside of the agent (and their cognitive system), and makes it inaccessible by the agent's consciousness (or subconscious cognitive systems). This has the unfortunate result that the agent (and his or her subconscious cognitive system) is unable to track the reliability of the processes in question. To try to address this, an internalist version of reliabilism (similar to that developed by Steup [2004, 2013, 2016]) might suggest that the system *can* try to track the *evidence* that a process tends to produce true beliefs, and that this can provide guidance internal to the system, while still retaining the idea that the ultimate value of justification is found in the external truth relationship and the frequency with which it obtains. However, as I have argued throughout this dissertation, it is the satisfaction of our goals that makes our cognitive efforts valuable and successful (when they are), and the tendency of a process to produce beliefs and other states that cohere well with each other, with the outputs from other cognitive processes, and with the goal states, is the relationship that is of central epistemic importance to us. Thus, my theory goes so far as to hold that *both* the evidence *and* the valuable thing evinced are internal to the agent's cognitive system.

> *Desideratum 4: The theory should be genuinely informative about, and make use of, the subconscious/subdoxastic states that constitute a sizable and important share of human cognitive processing.*

The specific mechanisms, contributions, roles and relationships operative at the subdoxastic level are very clearly in need of much more research, both by scientists and philosophers. Even a casual glance at a neuroscience or psychology journal will indicate that the scientific community is well aware of the significance of, and need for, more research at this level. But I think that the

philosophical community generally cannot make the same claim, and this is a serious problem. I believe one of the important contributions of this dissertation is to develop a framework that opens the door for the inclusion of these sub-conscious processes, ones that are not accessible by agents as traditionally conceived, in our philosophical theorizing. I will readily admit, however, that this is one area of my theory that will require considerably more work and research in the future to fully illuminate the epistemic and cognitive working of the brain at the subdoxastic level. I do take myself to have provided arguments that it can, and should, be done, as well as initial ideas about how we might approach it. The development of my theory of justification was inspired by neuroscientific and cognitive scientific theories about how neuronal synaptic weights are increased and decreased in response to frequent co-firing, and as result, the reduction from my version of epistemic justification to neural network should require less "translation". As such, I take it to be clear that this further research is warranted. After all, if philosophy is to be taken seriously, it must make more of an effort to engage with the scientific work being done and we must be willing to adapt our theories in response.

The approach of epistemologists, understandably, has been tied heavily to the currently accepted views in philosophy of mind. Epistemology has been seen as a project of analyzing central epistemic concepts like knowledge and justification, and has proceeded by establishing prototypical classes of, for example, justified and unjustified beliefs, then trying to identify what is common to each in the hope of identifying necessary and sufficient conditions. Once these classes have been discovered by philosophical methods, it was then generally idealized that epistemologists would hand these criteria and theories over to neuroscientists to work out the "non-philosophical" (and non-epistemic) details, and figure out what physical realizers or other corresponds to the different epistemic mental states. This approach makes a great deal of sense

on a functionalist or other non-reductive physicalist approach to the nature of the mind, but we might wonder if our aims would be better served by refraining from this top-down approach and instead paying more attention to the physical mechanisms and theories discovered so far by our physical sciences and using these to shape our epistemic theorizing. This has the potential to help avoid the gap between epistemic theorizing and neuroscientific research that has been problematic since neuroscience and other fields have been producing empirical results. For example, many of our best contemporary scientific theories about the nature of the mind make use of connectionist paradigms, drawing on the highly connected, networked nature of the brain that we have been discovering. Many people also think that dynamical system techniques and models have a lot to offer in understanding the functioning of the brain. These approaches and others seem to provide promising routes of inquiry into traditional philosophical topics (like the nature of justification), and philosophical theorizing should make the effort to take these opportunities seriously. I have tried to do so in this project.

Desideratum 5: The theory should be pragmatically oriented and constrained.

As we saw earlier, I am by no means the first person to worry about the presumption that aiming at truth is central to epistemology. Given the concerns raised by myself and others, it would be both interesting and advantageous if a theory of justification could be found that tells a satisfactory story about how our how our cognitive processes should form and update our beliefs, and fits our intuitions across a wide range of cases involving perception, memory, introspection, reasoning and testimony. By focusing on pragmatic considerations, endo-reliabilism does just this. The justificatory status of beliefs is not tied to truth or truth-conduciveness in any way, but

instead to the coherence between the processes' outputs and their coherence to goal states. As a result, the theory developed here is thoroughly pragmatic, and meets this *desideratum.*

> *Desideratum 6: if possible, the theory should have normative force or at least provide an error theory for why justification seems normative.*

This is an area where naturalistic theories traditionally do not fare well. Kim (1988) and others have criticized naturalistic theories on the grounds that they fail to deliver theories that are robustly normative. As we saw in the last chapter, however, we naturalists frequently accept a weaker notion of normativity, one closer to what is operative in scientific practice, and it is this standard that my theory aims to meet. After all, it is not the role of science to ask *why* we ought to want to cure cancer. But assuming that we do, science has a lot to tell us about how we should go about it and which methods are likely to be more successful and produce better results. Similarly, I will leave it to the ethicists and meta-ethicists to identify why we have the particular goals that we do, what makes them valuable, and whether some are better than others. But *assuming* that we want to accomplish our goals, then my naturalistic theory of epistemic justification has a lot to tell us about what our cognitive systems *ought* to do, and what training and practice we *should* subject ourselves to, etc. Some processes are more endo-reliable than others, and given our goals, it is *better for us* to use processes that are endo-reliable. I find this to already offer sufficient normativity for our purposes, and to help explain our everyday uses of, and intuitions about, these normative terms.

Part of the problem with assessing whether a theory is genuinely normative is that often philosophers are unclear about what is meant by the term *normative*, without realizing it. There is frequent confusion both at the individual- and discipline-level. I think it likely that there are

actually a plurality of concepts, understandings, and theories about what it is for something to be normative, and so I am not overly concerned about not adequately meeting the requirements for any particular understanding of normativity. Still, the last chapter included discussion that showed that endo-reliabilism still allows for adequate normativity, and helps point the way for how to improve our belief adoption and modification.

Overall then, I think it is fair to say that endo-reliabilism meets the *desiderata* established previously. Given that no other theory of which I am currently aware meets these *desiderata* as well as endo-reliabilism, then to the extent that they have been correctly identified, this gives us good reason to adopt endo-reliabilism as our theory of justification.

6.2 Value beyond the Desiderata

Of course, not everyone will find themselves accepting all of these *desiderata*. In that case, I think that this project still has worth as an exploration of the hypothetical situation where one *did* endorse all the *desiderata*. I find it interesting, however, that in discussions of this project with various colleagues and peers at conferences and other places where philosophers congregate, quite a few people stated that they agree with all of the *desiderata*, except for one of them. Which one, however, varied widely. And if each of the *desiderata* strikes a reasonable number of philosophers as plausible, then it seems a worthwhile investigation to examine what happens when they are all grouped together and used to steer the development of a theory of justification.

Still, many readers will find many of the claims and *desiderata* discussed above to be implausible. Some will lament that my theory does not meet their standards for *X*, or does not produce desired result *Y*. Like any good skeptic, I take myself to have at least exposed some places in contemporary epistemology where things are being assumed that perhaps should not be.

Perhaps the external world does not exist, or is not the way that we assume it to be. Perhaps our folk notion of consciousness does not correspond to a real thing, or at least conscious access is not the essential epistemic boundary is often treated as being. Perhaps the unconscious workings of our brains are just as important to justification as our conscious processes. Perhaps truth is not accessible to us as a part of our epistemic theorizing, and is not actually the proper goal of justification as has been assumed by so many for so long. Perhaps our notions of epistemic normativity are confused, mistaken, and misguided. The skeptic works by pointing to things that make us uncomfortable, but that upon dialogue and reflection, can often be seen to have some merit. Unlike most skeptics that challenge our assumptions and then move on, I have also tried to offer a solution. Specifically, I have tried to identify and develop an alternative theory that respects the limitations suggested by the skeptical approach, but still delivers as much as we can get from within that framework.

I believe this is all in keeping with the naturalist spirit. In many ways, it is the essence of scientific inquiry to try generating different theories by replacing different variables with different values and see what happens. Sometimes the resulting theory is inferior to what was previously available. But sometimes it is an improvement, and the only way to find out is to conduct the experiment and observe the results. Scientific history has also shown that we should be willing to challenge our assumptions, even the ones we are most confident in. While perhaps cliché, only by being willing to question what was allegedly obvious were the discoveries of Copernicus, Newton, Darwin, Einstein, and so many others possible. Similarly, history shows us that epistemology makes progress through dialogue with the spirit of skepticism, and so it would be a grave misstep to simply gag the skeptic and pretend the skeptical standpoint does not exist or does not warrant serious, open-minded investigation and exploration.

6.3 New Directions

By this point, it should be clear what I take to be the significance of endo-reliabilism for epistemology and the naturalistic project. There are some additional, perhaps surprising, areas where I think that the approach I have argued for might be useful, and so call out for further research.

First, consider our efforts to develop artificial intelligence. The nature of this project, where we, as the designers, are in an important sense *outside* of the systems we are developing, gives us more of exactly the kind of access to external assessments to which I criticize externalists for helping themselves. So, I admit that this is an area where traditional process reliabilism may prove quite helpful. A computer programmer may well wish to evaluate how successful a programmed subroutine is at consistently generating the results that she, with an experiential perspective outside of the computer system's limitations, thinks that it should. Of course, I still maintain that she does not have adequate access to the actual *objective* truth of the matters in question, but there are some important differences in this situation. Because humans are developing the systems in question, and are currently unable to incorporate all of our same cognitive abilities in the systems we design, our cognitive systems retain capabilities that the designed systems lack. For this reason, there is a sense in which the processes had by the designed system will be a proper subset of those possessed by our cognitive systems and our additional abilities afford us a perspective that exceeds that of the designed systems, in at least some ways, and place us in a position to access some information that is external and inaccessible to the designed system. As a result, we stand in a relationship to the computer system that allows *us* to employ a *more* externalist-style approach to assessing the reliability of

its processes. For example, if a researcher is developing a program for a robot to navigate an obstacle course, the researcher will be the one positioning the objects in the course, and so has access to information from before the robot is activated, as well as a clear set of expectations for what counts as correct results. The researcher is then able to check the actual results against this (more) external set of desired results and assess the reliability of the processes in question.[110]

However, if our interest is to develop functions and routines similar to our own, I think it likely that endo-reliabilism and the story it offers about the interrelations of internal processes, their level of reliability in cohering with each other, and their satisfaction of goal states, without any requirement of conscious access, could potentially prove quite helpful and informative. Endo-reliabilism locates all of the justificatory relata inside of the system, which allows it to be accessed and utilized by the system in question. This provides a model that AI development could potentially use for programming and developing self-assessment sub-routines and learning algorithms.

Second, I think it is possible that, with further research, endo-reliabilism could provide a helpful bridge between traditional analytic epistemology and feminist approaches to epistemology. My approach has been to try to develop a theory that merges many of what I consider to be the best insights from a range of philosophical work. Several of these insights are also frequently employed in feminist epistemological theorizing. For example, many feminist epistemologists are committed to both the naturalistic project and replacing the truth-orientation of epistemology with a pragmatic one (Anderson, 2015). Endo-reliabilism, by its very nature, is

[110] Of course, as humans we do something similar whenever we check our results with other people. This inter-subjective assessment and calibration is extremely valuable, but it must be remembered that, even if the other subjects are real and not computer simulation, we are all in similar epistemic situations, and so share roughly the same limitations (discussed earlier). As a result, no human has the sort of higher-level access that a computer programmer has relative to an AI system.

more amenable to the recognition and inclusion of a plurality of aims and social situations, and so may integrate better with feminist approaches than other theories that were not explicitly developed within a feminist framework.

Finally, my theory also has significant consequences for numerous other social issues. I think recent political results have shown just how intimate the link is between our epistemic practices and the pressing social issues we face. Overcoming confirmation bias, stereotype threat, and implicit bias are but a few of the many challenges that we all face, and that traditional epistemic theories are poorly equipped to handle. As we have seen, one of the most important features of my theory of justification is the inclusion of *subconscious* mental states and cognitive processes. Recognizing the important role they play in shaping and justifying our beliefs opens the door to better understand and address issues like individual and systematic racism, sexism, and other forms of discrimination, as well as a host of related topics in social epistemology.[111]

6.4 Takeaways and Summing Up

Overall, endo-reliabilism has some promising applications and some distinct advantages over both traditional coherentism and process reliabilism, while still being compatible with a naturalistic approach to epistemology. My theory of justification will fit particularly well with Quinean-style holism, which can be seen to still exert considerable influence in the epistemological research being performed by Paul Churchland (e.g., 2007, p. 107), Penelope Maddy (2005), Louise Antony (2004), and many others, and may facilitate a relatively

[111] Goldman (1993 and 1999, for example) has already explored some of the application of process reliabilism to social epistemological issues, but given the pronounced differences between our theories, it is likely that my approach would produce considerably different, and perhaps more useful, results. This would, of course, require further investigation.

straightforward reduction of justification to psychological and neuroscientific terms. There has been a gap between normative epistemological theorizing in philosophical practice and the biological investigation undertaken in neuroscience and psychology, and my project aims to bridge this gap by presenting a normative epistemic theory of justification that is reducible to (or at least realizable by) the kinds of biological mechanisms and interactions found at the neural level.

It also has some aspects that will be seen by many as disadvantages, especially its decoupling of justification from truth. However, rather than downplaying the challenge that skepticism poses, the view on offer investigates what epistemic system can be built within the confines of the skeptics worries. I am hopeful that, by making use of the tools of pragmatism, coherence, and scientific theories, the system constructed meets *most* of our normal, everyday intuitions and desires concerning justification and epistemic norms. Perhaps someone will succeed in a more optimistic approach, but at the least we should be aware of and explore an epistemic "worst case scenario".

At the same time, there are naturalists who are concerned that an externalist theory of justification cannot possibly yield the regulatory role of justification that we expect it to have. My internalist, naturalistic theory of justification resolves many of these worries, while also making naturalism more resilient to attack, since, if standard externalism is shown to be false (perhaps by the new evil demon problem, the generality problem, etc.), naturalism is better off if independent from it. Tying naturalistic epistemology to externalist theories of justification results in having all attacks against externalism also constitute attacks against naturalistic epistemology, and so if naturalized epistemology can be shown to be *also* compatible with a range of internalist

approaches to justification, it is thereby made at once more attractive to a broader philosophical audience and more philosophically resilient.

Whether the theory of justification I have argued for is “True” or not, I do not know. But hopefully I will have convinced my reader that it is the most coherent, theoretically virtuous, and pragmatic account that can be constructed given the present cognitive webs within which we find ourselves caught.

Bibliography

Anderson, Elizabeth, "Feminist Epistemology and Philosophy of Science", The Stanford Encyclopedia of Philosophy. Fall 2015 Edition, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/feminism-epistemology/>.

Antony, Louise. "A Naturalized Approach to the A Priori" in. *Epistemology*, edited by Ernest Sosa and Enrique Villanueva. 14 Vol. Boston, MA; Oxford, UK: Blackwell Publishing, 2004.

Armstrong, David., *A Theory of Universals*, Cambridge: Cambridge University Press, 1978.

---. *What Is a Law of Nature?*, Cambridge: Cambridge University Press, 1983.

---. "What Makes Induction Rational?", *Dialogue*, 30, (1991), 503–511.

---. "The Identification Problem and the Inference Problem", *Philosophy and Phenomenological Research*, 53, (1993), 421–422.

Audi, Robert. *Epistemology: A Contemporary Introduction (Routledge Contemporary Introductions to Philosophy).* 2nd ed., Routledge, 2002.

Bermúdez , José Luis. *Cognitive Science: An Introduction to the Science of the Mind.* New York: Cambridge University Press, 2010.

Bishop, Michael. "Why the Generality Problem is Everybody's Problem", *Philosophical Studies*, 151, (2010), 285–298.

Bishop, Michael and Trout, J. D. *Epistemology and the Psychology of Human Judgment.* New York: Oxford University Press, 2005.

Block, Ned. "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy*, 10, (1986), 615–678.

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

BonJour, Laurence, and Ernest Sosa. *Epistemic Justification: Internalism Vs. Externalism, Foundations Vs. Virtues (Great Debates in Philosophy)* . Wiley-Blackwell, 2003.

Bostrom, Nick. "Are you Living in a Computer Simulation?", *Philosophical Quarterly* , Vol. 53, No. 211 (2003), 243-255.

Bourget, David and Chalmers, David J. "What do philosophers believe?", *Philosophical Studies,* 170 (3), (2014), 465-500.

Bovens, Luc, and Stephan Hartmann. *Bayesian Epistemology*. Oxford: Clarendon, 2003.

Carruthers, Peter. *The Opacity of Mind: An Integrative Theory of Self-knowledge.* Oxford, Oxford University Press, 2011.

Cherniak, Christopher. "Computational Complexity and the Universal Acceptance of Logic", *The Journal of Philosophy* (1984). Reprinted in *Naturalizing Epistemology*, 2nd ed. edited by Hilary Kornblith, Cambridge, MA: MIT Press, 1994.

---. *Minimal Rationality*. Cambridge, MA.: MIT Press, 1986.

Chisholm, Roderick. "The Indispensability of Internal Justification." *Synthese* 74:3, (1988), 285-296.

Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy*, Vol.78, No.2. (Feb.,1981), 67-90.

---."Folk Psychology and the Explanation of Human Behavior". *Philosophical Perspectives*, Vol.3, Philosophy of Mind and Action Theory. (1989), 225-241.

---.*The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press, 1995.

---. *Neurophilosophy at Work.* Cambridge: Cambridge University Press, 2007.

---.*Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals.* Cambridge: MIT Press, 2012.

Churchland, Patricia Smith. "Epistemology in the Age of Neuroscience." *The Journal of Philosophy*, Vol. 84, No. 10, Eighty-Fourth Annual Meeting American Philosophical Association, Eastern Division (Oct., 1987), 544-553.

Comesaña, Juan. "The Diagonal and the Demon", *Philosophical Studies*, 110, (2002), 249–266.

---."Evidentialist Reliabilism", *Noûs*, 94, (2010), 571–601.

Conee, Earl, and Richard Feldman. "Evidentialism", *Philosophical Studies* 48, (1985), 15-34.

---. *Evidentialism: Essays in Epistemology*. New York: Oxford, 2004.

Cosmides, Leda and John Tooby, "Cognitive Adaptions for Social Exchange". In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, edited by Jerome Barkow, Leda.Cosmides, and John Tooby, 163–228, New York: Oxford University Press, 1992.

Cottrell, Gary. “Extracting Features from Faces using Compression Networks: Face, Identity, Emotion and Gender Recognition using Holons.” *Connectionist Models: Proceedings of the 1990 Summer School*, (1990), 328-337.

Cottrell, Gary, Panqu Wang , and Isabel Gauthier. “Experience matters: Modeling the Relationship between Face and Object Recognition.” In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2014.

Da Costa, Newton and Steven French. *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford: Oxford University Press, 2003.

de Gelder, Beatrice, Tamietto M., van Boxtel G., Goebel R., Sahraie A., van den Stock J., Stienen BM., Weiskrantz L., Pegna A., “Intact Navigation Skills After Bilateral Loss of Striate Cortex.” Current Biology Volume 18, Issue 24, (Dec., 2008), 1128–1129.

Dennett, Daniel C. “Intentional Systems”. *The Journal of Philosophy*, Vol.68, No.4. (Feb.25,1971), 87-106.

---.“True Believers: The Intentional Strategy and Why it Works.” Scientific Explanations (1981) in *Philosophy of Mind: Classical and Contemporary Readings*. Edited by David J. Chalmers, 556-568. Oxford: Oxford University Press, 2002.

---.“Real Patterns.” *The Journal of Philosophy*, Vol.88, No.1. (Jan.,1991), 27-51.

---. *Consciousness Explained***,** London: Allen Lane, 1991.

---.“Where am I?” in *Brainstorms: Philosophical Essays on Mind and Psychology*, 310-324, MIT Press, 1981.

Dretske, Fred. “Conclusive Reasons.” *Australasian Journal of Philosophy* 49, (1971), 1-22.

---.“Laws of Nature”, *Philosophy of Science*, 44 (1977), 248–268.

---. *Knowledge and the Flow of Information*, Cambridge, MA: MIT/Bradford Press, 1981.

Feldman, Richard. “Reliability and Justification,” *The Monist* 68:2, (1985), 159-74.

Fodor, Jerry. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford, 1987.

Gertler. Brie. *Self-Knowledge (New Problems of Philosophy).* New York: Routledge, 2011

Gettier, Edmund. "Is Justified True Belief Knowledge?" *Analysis* 23, (1963), 121-123.

Giere, Ronald. “How Models Are Used to Represent Reality.” *Philosophy of Science*, Vol. 71, (December 2004), 742-752.

Gigerenzer, Gerd. *Simple Heuristics That Make Us Smart.* New York: Oxford University Press, 1999.

Gigerenzer, Gerd, Peter Todd, and the ABC Research Group. *Ecological Rationality: Intelligence in the World*, Oxford University Press; 2012.

Goldman, Alvin I. "A Causal Theory of Knowing." *The Journal of Philosophy* 64, no. 12 (Jun. 22, 1967), 357–372.

---. "Discrimination and Perceptual Knowledge," *Journal of Philosophy* 73 (Nov. 1976), 771-791.

---. "What is Justified Belief?" in *Justification and Knowledge*, edited by George Pappas. Boston: D. Reidel, 1979. Reprinted in in *Naturalizing Epistemology*, 2nd Edition, edited by Hilary Kornblith, 105-130. Cambridge: MIT Press, 1994.

---. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press, 1986.

---. "Epistemic Folkways and Scientific Epistemology". *Philosophical Issues*, 3 (1993), 271-285.

---. *Knowledge in a Social World*. Oxford: Clarendon Press, 1999.

---. *Pathways to Knowledge: Private and Public*. New York: Oxford University Press, 2002.

---. *Readings in Philosophy and Cognitive Science*. Cambridge, MA: MIT Press, 1993.

---. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press, 2006.

---. "Toward a Synthesis of Reliabilism and Evidentialism", in *Evidentialism and Its Discontents*, edited by T. Dougherty, 254–290. New York: Oxford University Press, 2011.

---. *Reliabilism and Contemporary Epistemology: Essays*. New York: Oxford University Press, 2012.

Goldman, Alvin and Bob Beddor. "Reliabilist Epistemology", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/reliabilism/>.

Grundmann, Thomas. "Reliabilism and the Problem of Defeaters." *Grazer Philosophische Studien*, 79 (1), (2009), 65-76.

Haught, James. *Holy Horrors: An Illustrated History of Religious Murder and Madness*. Amherst, NY:Prometheus Books, 1990.

Henderson, David and Terry Horgan, "The Ins and Outs of Transglobal Reliabilism" in *Internalism and Externalism in Semantics and Epistemology*, edited by Sanford Goldberg, 100-130. New York: Oxford University Press, 2007.

Henrich, J., Heine, S., and Norenzayan, A. "The WEIRDest people in the world?", *Behavioral and Brain Sciences*, 33, (2011), 61–135.

Hofstadter, Douglas and Daniel Dennett, *The Mind's I*, New York: Bantam, 1982.

Huemer, Michael. *Skepticism and the Veil of Perception.* Lanham, MD: Rowman & Littlefield Publishers, 2001.

---. *Epistemology: Contemporary Readings*, edited by Michael Huemer. London and New York: Routledge, 2002.

Kahnemen, Daniel. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, 2011.

Kant, Immanuel. *Critique of Pure Reason*. (CPR) Trans. Paul Guyer and Allen Wood, Cambridge: Cambridge University Press, 1998.

Kim, Jaegwon. "What is 'Naturalized Epistemology?'". *Philosophical Perspectives*, 2, (1988), 381-405.

---.*Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press, 1993.

---. *Philosophy of Mind*, Cambridge: Westview Press, 2006.

Kitcher, Philip. "The Naturalists Return**,**" *Philosophical Review***,** 101, (1992), 53-114.

Klein, Peter, and Ted Warfield. "What Price Coherence?," *Analysis*, 54, (1994), 129–132.

—— "No Help for the Coherentist", Analysis, 56, (1996), 118–121.

Kornblith, Hilary. "Beyond Foundationalism and the Coherence Theory," *Journal of Philosophy*, 77 (1980), 597–61.

---. "How Internal Can You Get?" *Synthese*, Volume 74, Issue 3, (March 1988), 313-327.

---. "Epistemic Normativity", *Synthese*, Volume 94, (1993), 357–76.

---. *Naturalizing Epistemology*, 2nd Edition, Cambridge: MIT Press, 1994.

---. *Epistemology: Internalism and Externalism*. 2 Vol. Malden, MA: Blackwell Publishers, 2001.

---. *Knowledge and its Place in Nature*. Oxford: Clarendon Press, 2002.

Kvanvig, Jonathan. *The Value of Knowledge and the Pursuit of Understanding.* Cambridge: Cambridge University Press, 2003.

Laakso, Aarre and Garrison Cotrell, "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems." *Philosophical Psychology*, 13, (2000), 47–76.

Laudan, Larry. "A Confutation of Convergent Realism", *Philosophy of Science*, Vol. 48, No. 1, (Mar. 1981), 19-49

Lehrer, Keith. *Theory of Knowledge.* Boulder, Colorado: Westview, 1990.

Leibniz G.W., *Monadology* (1714) in *G.W. Leibniz's Monadology* edited by Nicholas Rescher, Abingdon, Oxon: Routledge, 2014.

Littlejohn, Clayton. *Justification and the Truth-Connection*. Cambridge: Cambridge University Press, 2012.

Lyons, Jack. *Perception and Basic Beliefs: Zombies, Modules, and the Problem of the External World.* Oxford: Oxford University Press, 2009.

Maddy, Penelope. "Three Forms of Naturalism." in *The Oxford Handbook of Philosophy of Mathematics and Logic,* edited by Stewart Shapiro, Oxford University Press, 2005.

Maffie, James. "Naturalism and the Normativity of Epistemology", *Philosophical Studies*, 59, (1990), 333–349.

Marcus, Gary. Kluge: *The Haphazard Construction of the Human Mind*. New York: Houghton Mifflin Co, 2008.

McDowell, John Henry. *Mind and World*. Cambridge, MA: Harvard University Press, 1994.

Millikan, Ruth.*Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press, 1984..

Mitchell, Sandra. "Dimensions of Scientific Law." *Philosophy of Science*, Vol. 67 (2), (2000), 242-265.

Nisbett, Richard, and Timothy Wilson. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review*, 84, (1977), 231-259.

Nozick, Robert. "Knowledge and Skepticism." in *Philosophical Explanations*. Cambridge, MA: Harvard University Press, 1981.

Okruhlik, Kathleen. “Gender and the Biological Sciences” in *Biology and Society, Canadian Journal of Philosophy*, Supplementary volume 20 (1984), 21-42. Reprinted in *Philosophy of Science: The Central Issues*, edited by Curd and Cover, W.W. Norton & Co., Inc, 1998.

Pollock, John L., and Joseph Cruz. *Contemporary Theories of Knowledge*. 2nd ed., Lanham, MD: Rowman & Littlefield Publishers, 1999.

---. “The Chimerical Appeal of Epistemic Externalism", in *The Externalist Challenge: Studies on Cognition and Intentionality*, edited by Richard Schantz, New York: de Gruyter, 2004.

Prinz, Jesse. ‘Empirical Philosophy and Experimental Philosophy’ in *Experimental Philosophy*, edited by J. Knobe and S. Nichols. Oxford: Oxford University Press, 2008.

---. *The Conscious Brain: How Attention Engenders Experience.* Oxford, Oxford University Press, 2012.

Pritchard, Duncan. *Epistemic Luck*. Oxford: Oxford University Press, 2005.

Psillos, Stathis. *Scientific Realism :How Science Tracks Truth*. New York: Routledge, 1999.

Putnam, Hilary. *Reason, Truth, and History*. Cambridge: Cambridge University Press, 1981.

Quine, W. V.O. “Two Dogmas of Empiricism”, *The Philosophical Review* 60, (1951), 20–43.

---. *Word and Object*. Cambridge, MA: MIT Press, 1960.

---. “Epistemology Naturalized”, *Ontological Relativity and Other Essays,* 1969, reprinted in *Naturalizing Epistemology*, 2nd ed, edited by Hilary Kornblith, Cambridge, MA: MIT Press, 1994.

---. *From a Logical Point of View*. 2d ed. Cambridge: Harvard University, 1980.

---. “Reply to Morton White,” in *The Philosophy of W. V. Quine*, edited by Lewis Edwin Hahn and Paul Arthur Schilpp, 663-665, LaSalle, IL: Open Court, 1986.

---. *Pursuit of Truth*. Cambridge, MA: Harvard University Press, 1990.

Quine, W.V.O., and J.S. Ullian. *The Web of Belief.* New York: Random House, 1970.

Rain Man. Directed by Barry Levinson. Santa Monica, CA: MGM Home Entertainment, 1988.

Reed, Geoffrey, Kemeny, Margaret, Taylor, Shelley E.;Wang, Hui-Ying J.;Visscher, Barbara R., “Realistic Acceptance as a Predictor of Decreased Survival Time in Gay Men with AIDS”, *Health Psychology*, Vol 13-4, (Jul 1994), 299-307.

Rupert, Robert. “The Best Test Theory of Extension: First Principle(s),” *Mind & Language*, 14, (1999), 321–355.

---. *Cognitive Systems and the Extended Mind.* New York: Oxford University Press, 2009.

Russell, Bertrand. *The Problems of Philosophy*, Oxford University Press, Oxford, 1912, reprinted 1978.

Schantz, Richard. *The Externalist Challenge*. New York: Walter de Gruyter, 2004.

Sosa, Ernest. “How to Defeat Opposition to Moore.” *Philosophical Perspectives*, 13, Epistemology, (1999), 141-54.

---. “Skepticism and Contextualism.” *Philosophical Issues*, 10, Skepticism, (2000), 1-18.

Steup, Matthias. *An Introduction to Contemporary Epistemology*. Upper Saddle River, NJ: Prentice-Hall, 1996.

---. “Internalist Reliabilism”. *Philosophical Issues* 14-1, (2004), 403–425.

---. “Does Phenomenal Conservatism Solve Internalism’s Dilemma?” in *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism*, edited by Chris Tucker, 135-153. New York: Oxford University Press, 2013.

---. “Destructive Defeat and Justificational Force: The Dialectic of Dogmatism, Conservatism, and Meta-Evidentialism”, *Synthese*, (August, 2016), 1-27.

Stich, Stephen. *The Fragmentation of Reason : Preface to a Pragmatic Theory of Cognitive Evaluation.* Cambridge, MA: MIT Press, 1990.

Taylor, S.E., Kemeny, M.E., Aspinwall, L.G., Schneider, S.G., Rodriguez, R, Herbert, M. “Optimism, Coping, Psychological Distress, and High-Risk Sexual Behavior Among Men at Risk for Acquired Immunodeficiency Syndrome (AIDS)”, *Journal of Personality and Social Psychology*, 63-3, (1992), 460-473.

Thagard, Paul. *Conceptual Revolutions*. Princeton: Princeton University Press, 1992.

Tooley, Michael. “The Nature of Laws”, *Canadian Journal of Philosophy*, 7, (1977) 667–698.

---. *Causation*, Oxford: Clarendon Press, 1987.

Tversky, Amos and Daniel Kahneman. “Judgment Under Uncertainty: Heuristics and Biases.” *Science*, 185, (1974), 1124-1131.

van Fraassen, Bas, *Laws and Symmetry*. Oxford: Clarendon Press, 1989.

Wason, P. C. "Reasoning". In *New Horizons in Psychology*, edited by B. M. Foss, Harmondsworth: Penguin. 1966.

Weinberg, Jonathan. “What's Epistemology For?: The Case for Neopragmatism in Normative Metaepistemology.” In *Epistemology Futures*, edited by S. Hetherington, 26-47, Oxford: Oxford University Press, 2006.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich, “Normativity and Epistemic Intuitions.” *Philosophical Topics*, 29-1&2, (2001), 429–459.

Williamson, Timothy. *Knowledge and its Limits*. Oxford: Oxford University Press, 2000.

---. *The Philosophy of Philosophy*, Malden, MA and Oxford: Blackwell, 2007.

Wilson, Margaret (ed.). *The Essential Descartes*, New York: Mentor, 1969.

Appendix

Here, I want to offer an idea for a more precise and formal way of understanding the role that theoretical virtues play in the theory evaluation process. The key to this suggestion is the use of hyper-dimensional phase spaces. Physicists have long been aware of the utility of these tools, but work by Sandra Mitchell (2000) on scientific laws and Paul Churchland (1995, 2007, and 2012) on neurological activation vectors, as well as the recent trend in cognitive science to employ dynamical systems models, has shown the power and usefulness of hyper-dimensional phase space in understanding and modeling other complex possibility spaces. Let us examine how this approach can be applied to theory evaluation.

Just like we plot a graph by assigning one variable to the x-axis, one to the y-axis, and another to the z-axis, we can assign an axis to represent the possible values of a particular theoretical virtue. For example, if we assign coherence to be represented by the x-axis, a theory consisting of the two statements "A" and "~A" would have its coherence value plotted exactly at the origin, since its coherence value is zero. On the other hand, a theory with only the statement "A" is perfectly coherent, and so would be assigned the maximum value (let's follow the standards in probability theory notation and assign the maximum value 1.0). All other theories will fall somewhere on this range between zero and one.[112]

We can then assign the y-axis to represent another theoretical virtue such as predictive power, the z-axis to represent simplicity, and so on, assigning one dimension to each virtue.

[112] Bovens and Hartmann, as we've discussed, provide a formal and precise way to evaluate the coherence of any information set.

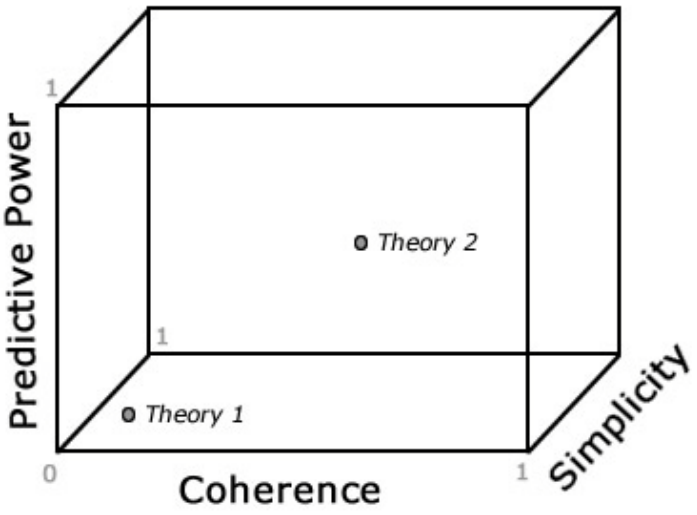

Figure 4 Hyper-dimensional theoretical virtue phase space.

Of course, the list of theoretical virtues includes more than three items, and so there will be additional dimensions needed. It is at this point, to avoid trying to "picture" an n-dimensional phase space, that we may wish to switch to vector representation instead. By assigning a theory's value for each theoretical virtue to a place in the ordered *n*-tuple, we are left with a precise mathematical representation of the theory's virtues. So, if we (arbitrarily) order the vector <coherence, predictive power, simplicity>, then on the above picture, Theory 1 might be represented as <0.10, 0.17, 0.31> while Theory 2 is < 0.64, 0.57, 0.42>. We can then subtract the values for Theory 1 from Theory 2, yielding <0.54, 0.40, 0.11>, and get a clear sense of which theoretical virtues would be increased or decreased, and to what extent. This may be especially helpful in cases where one theory is more theoretically virtuous in some ways, but less in others, as this will allow numerical comparison between the changes. Indeed, if we could make the value assignment, normalization, and weighting of the different components accurate enough, it would even be possible to sum all of the different vector values to yield a "total virtue score" for

each theory, and a “total virtue change score” that indicates the net improvement or loss to be had from moving from one theory to the other.

While this would require more work to flesh out, I think this approach clearly allows for a better assessment of the trade-offs and improvements made by competing theories, and a more precise method of comparison between them. It offers both an intuitive and powerful tool for deciding between theories, and understanding the strengths and weaknesses of each in a more explicit way.