BELIEF-DESIRE COHERENCE

by

Stephen Petersen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2003

Doctoral Committee:

Associate Professor Eric Lormand, Chair
Associate Professor James Joyce
Assistant Professor Thad Polk
Assistant Professor Jessica Wilson

ACKNOWLEDGEMENTS

Tradition compels me to write dissertation acknowledgements that are long, effusive, and unprofessional. Fortunately for me, I heartily endorse that tradition.

First, of course, I want to thank my chair, Eric Lormand. If our marathon advising meetings (of up to six hours) hadn't guaranteed me many times the usual amount of mentoring a graduate student receives, then the sheer frequency of our sessions would have. I've learned an enormous amount of philosophy from him in that time, and though we disagree on some key points, it's against a wide background of his views that I have simply absorbed. More importantly, I have also learned a great deal about *how* to do philosophy from Eric. In summary: over the years his influence on my philosophical development has run very deep. Besides that, he even came to my improv shows, and copied Muppet tapes for me.

Next I want to thank Jessica Wilson. She stepped in a year and a half ago as my second advisor, to advise me on a field that (she claims) is not her specialty. Her help was just what I needed, and I don't think I could have finished without her. She is a natural mentor—she has a talent for mixing, in ideal proportions, genuine encouragement and enthusiasm with careful and constructive criticism.

Jim Joyce agreed late in the game to be my third committee member, but already by then he had helped me a great deal both philosophically and professionally, just in his role as general department *mensch*. And Thad Polk, my cognate committee member from cognitive psychology, rose above the call of what's too often a mere duty of formality.

He cheerfully tolerated my philosopher's take on cognitive science, and even pointed me toward some of the real stuff.

Many other professors expended a great deal of time and effort on my education; just about each member of the department has helped me at one point or another. I'd especially like to thank Ian Proops for a wonderful working relationship during my pre-candidacy, and continued moral support throughout my years here. David Hills and Jason Stanley also gave me much of their time and wisdom in my pre-candidacy days. Rich Thomason educated me about artificial intelligence where he could, and I loved debating philosophy with Dick Alexander over in the biology department. Louis Loeb and David Velleman served as model philosophy teachers for me.

The philosophy department staff were always there for me, and I'm especially grateful to Sue London, Linda Shultes, and Michele Bushaw Smyk. Among other things, they all helped me get the full benefit of the fantastic funding package that the philosophy department provided. I was always proud to tell the graduate student union here how generous the philosophy department is.

On the next rung of the mentoring ladder are my fellow graduate students. I owe a special debt to my seniors in the program, notably Jim Bell, Karen Bennett, Jeanine Diller, Jeff Kasser, Evan Kirchhoff, Katie McShane, Angela Napili, Gerhard Nuffer, Greg Sax, Laura Bugge Schroeter, Nishi Shah, and Kevin Toh. And though of course playing at friendship and philosophical apprenticeship with Robert Mabrito was just a ruse toward foiling his diabolical plans, it turned out especially edifying. My cohort had great influence on me in both senses of 'great'—especially Stephen Martin and Blain Neufeld. Among my "juniors" I have had friends and philosophical comrades like Steve Daskal, Remy Debes, David Dick, Alexa Forrester, Soraya Gollop, Charles Goodman, Liz Goodnick, Robert Gressis, Alex Hughes, Rob Kar, Hanna Kim, Matt Pugsley, Justin Shubow, Paul Sludds,

Tim Sundell, and Gabriel Zamosc-Regueros.

The last link of the university food chain are my many students, from whom I always learned a lot, and to whom I'm grateful.

Then there are just plain *friends*, who deserve more credit for my completing this dissertation than they likely know. *Tilt*, the comic improv group I founded and "artistically direct",[1] was especially crucial to my survival; no matter how down I got, they made me laugh and got me to *play*, week after week. I grew especially close to Robert Gressis, Steve Kime, and Jenni Pickett through the years—thanks for those crucial buckets of fish. Another hobby was *Project X*. (All its members have been named, so no need to repeat them here, but extra thanks are in order.) Hillary Holloway, one of my best friends for fifteen years now, had the good grace to share five of those in Ann Arbor with me while she got her own PhD. She was especially supportive during those worst of times, in 2001 and 2002. So were, alternately, Jenifer Clark and Hanna Kim. There are many other friends who made Ann Arbor a better place; I'll put them in a smaller-type footnote so that I don't exceed an embarrassing three pages.[2] Then there are the many friends from other times, or walks of life, or parts of the world.[3] I'm sure I've tragically forgotten some. It's okay; I suspect that, to paraphrase the oracle, they know who they are.

Thanks also, of course, to my family. And that reminds me: thanks also to my therapists. The discovery of therapy was an unintended side-effect of the dissertation process.

Finally, I'd like to thank the various telemarketers I've sued. After I ran out of departmental support, damages from their illegal activities funded my final year.

---

[1] The scare quotes are important.

[2] Tim Athan, Amy Burke, Anna Chen, Susan Chimonas, Jon Colman, Chris Consilvio, Alex DesForges, Eric Dirnbach, Anne Duroe, Rob Gray, Ben Hansen, Carmen Higginbotham, Jessica Hughes, Hana Ishizuka, Mark Krasberg, Kimberly Labut, Ji-Young Lee, Kevin Pimentel, Alix Schwartz, Kerry Sheldon, Melanie Sonnenborn, Becca Treptow, Mary Wagner, Mary Ann Wiehe, Wild Swan Theater, and Mina Yoo.

[3] Most notably: John Ackermann, Andy Brownstein, Evan Cohen, Heather Cross, Bill Egbert, Brad Elliott, Alison Gima, Steve Meyers, Linda Perlstein, Eiko Sakai, Andy Sernovitz, Sharon Sidlow, Elijah Siegler, Aaron Thieme, and Al Weiss.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **ii**

LIST OF FIGURES . . . . . **vii**

CHAPTER

I. INTERNAL PRAGMATIC EPISTEMOLOGY 1

1.1 Creatures and functions 2
1.2 Functions and internal epistemology 4
1.2.1 Internal functions 5
1.2.2 Epistemic norms and guidance 7
1.2.3 Guidance and epistemic agency 11
1.2.4 Guidance and internal functions 13
1.2.5 A concrete example 16
1.3 Internal epistemology and learning 18
1.3.1 Learning and intelligence 24
1.3.2 Learning and induction 26
1.4 The normative status of the standard 29

II. WISHFUL THINKING AND COHERENCE EPISTEMOLOGY . . 31

2.1 Wishful thinking 31
2.1.1 Epistemic relativism and "anything goes" epistemology 32
2.1.2 The wishful thinking objection 34
2.1.3 Wishful thinking as rational 35
2.1.4 Wishful thinking as irrational again 41
2.1.5 Thinkful wishing 44
2.1.6 A way out: spontaneous thoughts 46
2.1.7 External constraints 48
2.2 Coherence 52
2.2.1 Rational desires 52
2.2.2 Intelligence and rich aims 55
2.2.3 Defaults and foundherence 59
2.2.4 Coherence and humans 62

III. COMPUTATIONAL EPISTEMOLOGY 64

3.1 Belief-desire coherence (BDC) 64
3.1.1 Specifications 65
3.1.2 Machine learning 67
3.1.3 Formal coherence 69
3.1.4 The belief-desire coherence algorithm 71
3.1.5 The architecture, and a toy example 73
3.1.6 Next steps 77
3.2 Advantages and further issues 78
3.2.1 DECO, ECHO, and BDC 78
3.2.2 Explanations, contradictions, and BDC 83
3.2.3 Folk psychology and BDC 87
3.2.4 Emotions and BDC 90
3.2.5 Ethics and BDC 93
3.2.6 Conclusion 96

IV. TRUTH AND INTERNAL EPISTEMOLOGY 97

4.1 Can we aim at truth? 97
4.1.1 A warm-up argument 98
4.1.2 BDC, *a priori* beliefs, and the truth aim 100
4.1.3 Evidentialism 106
4.1.4 More general problems: content 109
4.1.5 More general problems: means to truth 112
4.1.6 Lormand's responsible searching 116
4.1.7 Dogmatism and truth 118
4.2 Should we aim at truth? 123
4.2.1 Skepticism and differences that matter 124
4.2.2 BDC and truth's value 132
4.2.3 The intrinsic value of truth 134
4.2.4 Truth conditions and truth's value 140
4.2.5 The instrumental value of truth 146
4.2.6 Meta-pragmatism 157
4.3 Concessions to truth 159
4.3.1 *A priori* beliefs and external aims 159
4.3.2 Pragmatic evidentialism 162

LIST OF FIGURES

Figure

3.1 A somewhat sophisticated, but unintelligent, creature. 75

3.2 A toy example of the BDCA. 76

CHAPTER I

INTERNAL PRAGMATIC EPISTEMOLOGY

Here are two great reasons for wanting to investigate intelligence: first, because we would like to build machines that are intelligent; second, because we would like to be more intelligent ourselves. Those who focus on the first reason tend to be in artificial intelligence, and those who focus on the second tend to be in normative epistemology. But people in both fields are interested in ways to make things smarter, and people in both spend at least part of their time trying to figure out just what it is to be smarter, anyway.

In the next three chapters I will develop my positive answer to this question: a computational, internal, pragmatic, coherence epistemology. I call the theory, somewhat inaccurately and incompletely, *belief-desire coherence* (or *BDC* for short).[1] I save the final chapter for those reluctant to give up a *truth*-based approach to normative epistemology.

Now I do not know exactly what constitutes intelligence, but in this chapter I do think I can give some interesting necessary conditions. For something to possess intelligence I claim first that it must be what I will call a *creature*. And second, for at least an interesting level of intelligence, that creature must be capable of *learning* in a certain way. Greater capacity for learning—on a measure to come later—correlates well with the intelligence

[1] Inaccurate because the coherence is not restricted to belief-level cognitions and desire-level conations, as we will see; incomplete because the name fails to include the view's computational nature, foundational impurities, and roots in an internal, learning approach to epistemology.

we are independently willing to attribute. Notably intelligent creatures (like pigeons, for example) can learn, adjusting their cognitive mechanisms toward improvement at some task. Especially intelligent creatures, like humans, can even learn how to *think* better: they can learn new ways to form beliefs and desires.

Learning of any kind requires, I will argue, an *internally available standard of correctness*. In particular, learning to think better requires an internal standard of correctness for thought. I will argue this internal standard for learning must take the form of matching cognitions to conations—something like what we might call subjective pragmatic success. As we will see later, there is a natural choice for an algorithm to compute according to this standard (given basic cognitions and conations), and a natural architecture to implement this algorithm that as it happens is like our own. Thus the standard and its potential implementation should be of interest both to those in cognitive science and normative epistemology.

1.1 Creatures and functions

Only some things are even in the running for the label "intelligent". Animals can be intelligent, and maybe someday machines could be, and maybe some sophisticated machines are now to some degree intelligent. But couches, rocks, and solar systems *definitely* are not; they are just not the kind of thing we would even consider potentially smart. I think that is because for a thing to count as intelligent, that thing must first be a *creature*. Of course this is not a sufficient condition; there are many things we would naturally call "creatures" that are not intelligent. Single-celled organisms, for example, do not seem to be intelligent in any recognizable sense, yet they are clearly creatures. I also would consider plants to be unintelligent creatures, though I am not sure whether that is squarely within ordinary usage.

At any rate here is a rough characterization of creatures that at least approximates the ordinary notion: to be a creature is to have at least one function that can be performed autonomously—that is, solely through the performance of sub-functions internal to the creature. I emphasize that this is rough, and in particular I am not sure about the "autonomous" clause. But I do not think this will matter in what follows; the only aspect I lean on philosophically is the *functional* approach.

The intuition is that creatures are things that have what we can loosely call *wants*—they have preferences for the world to be a certain way, and they do things in the world to achieve that way. Plants want their leaves aimed at the sun; robins want to stay warm in the winter; humans want better local theater. The first two can loosely be called "wants" because of functions those creatures possess and are capable of bringing about without assistance. By contrast the earth causes things nearby to fall toward it, but that is not even loosely a want of the earth's, for it is not a *function* of the earth to cause the world to be this way. And though knives have the function of cutting things, this is not even loosely a want for knives, because they have no sub-functions designed to bring about achievement of this function autonomously. To attribute creaturehood to something, we have to see it as having a function, *and* we have to see it as able to perform that function on its own through sub-functions. It is this latter requirement that allows us to see the thing as "trying" to achieve the basic function, and thus as having a want. Since 'want' is not really the right word, though, I will often call such functions of creatures *aims*.

Note that this definition does not require creatures to be members of some biological natural kind, and I take this to be an advantage. Suppose with impressive new technology we managed to reverse-engineer a nightingale, and then build a robot that mimicked its functional behavior in all important details. By this characterization, that robot can be just as creaturely as its prototype. To say otherwise is, to my mind, to be biologically chau-

vinistic. You might object that the robot bird cannot *autonomously* bring about its basic functions, since it required us to build it in the first place. But the biological nightingale had to be designed as well, by natural selection. (I will use 'design' in this inclusive sense throughout.) Things cannot be required to bring about their own existence, or set their own fundamental aims, in order to count as creatures.

Or you might instead object that by this standard, my thermostat is a creature. It has the function of keeping the room a certain temperature, which it performs autonomously via an internal function to correlate a simple thermometer (usually a coil of thin metal) and a switch for the heater. Here I bite the bullet: yes, my thermostat is a very primitive creature, according to my characterization. There is a reason that Daniel Dennett, for example, picks the thermostat for "as degenerate a case of an intentional system as could conceivably hold our attention for more than a moment."[2] We can attribute primitive intentional states to it that we cannot for, say, a radio. I think the line I have drawn for creatures explains why. It provides a lower limit for things toward which we can take what Dennett calls the "intentional stance". For similar reasons we can attribute primitive intentional states to the new vacuum cleaners that scoot about the floor on their own, and maybe for my email program's Bayesian spam filter, and so on.

1.2 Functions and internal epistemology

The simple notion of a creature already gives us purchase on internal approaches to epistemology. Before I show how, I should emphasize that taking an internal approach to epistemology need not make one an internal*ist*; there may be several worthwhile epistemic norms, some internal and some external.[3] I only wish to argue that one clear reason to

[2]Dennett (1981) p235.

[3]See for example Alston (1993) or DePaul (2001) for pluralistic views of epistemology. The core of DePaul's argument is short and sweet enough to present here: if knowledge is better than true belief, then there is some epistemic norm other than truth!

examine epistemic norms demands an approach like the one that has traditionally been called internal. So my proposal opposes external*ism* to the extent that externalism denies the existence of any interesting internal epistemic standard, but does not oppose external epistemic projects in general.

1.2.1 Internal functions

Entities gain functions by design, which is (roughly) a process of giving feedback toward a goal. The goal can be explicitly represented in the designers, as when primitive humans made knives out of rocks by chipping them. In effect they gave the shape of the rock feedback by encouraging some shapes and discouraging others, toward their goal of sharper rocks. The resulting knife thereby has the function of cutting things. Such notions of design depend on intentionality, but there are naturalistic approaches to functions that do not require the explicitly goal-directed feedback of an intelligent designer. The goal can also be implicit in the feedback, as for example when the environment gives feedback to genome types through natural selection. More precisely (but still roughly): we properly say the effect of a thing is a *function* when the explanation for the thing's existence depends in the right way on successful achievement of that effect by it or others of its ancestral line.[4] Achieving the effect gives positive feedback by encouraging the existence of that thing or its replicated progeny.

If feedback designs a functionally organized entity with enough sophistication, that entity can develop its own *internal* functions, and we have the makings of a creature. For example, I have mentioned the thermostat has an internal function to correlate a crude thermometer and a heater switch. Optimizing a function completely internal to the thermostat

[4]This etiological-selectional account of functions is based on the more precise work of Ruth Garrett Millikan, as in Millikan (1984). I think, though, that my account might also be amenable to the "homeostatic" characterization of functions, and perhaps to the "instrumental" characterization as well. For references see the helpful Manning (1998).

allows it to optimize the external function of regulating room temperature for which it was designed. Similarly, evolution designed hermit crabs with an internal function that correlates food-like splotches in their visual field with trajectories of motor-nerve stimulation to their pincers. This allows them to optimize their external function of getting food (which in turn allows them to optimize the function of passing on genes).[5] Similarly again, we humans may have been designed by evolution to have functions for gathering (mostly) true beliefs. These functions may be optimized via an *internal* function that correlates with true beliefs—such as, oh, say, maximizing some kind of belief-desire coherence.

In general let's say that "internal" functions are such that you would have to change the thing's intrinsic properties to change its success in achieving that purposed effect; "external" functions are such that you could alter success in achieving the relevant effect just by changing some relational properties of the thing.

In each of the examples above, there is only a contingent connection between the success of the internal function and the success of the external one. Someone might be holding a lighter under the thermostat's thermometer, or it could be wrapped in an ice pack. The switch could be connected to a broken heater, or not connected to a heater at all. If so the thermostat could still optimize its internal function—correlating the thermometer with a switch—and yet not optimize its external function of keeping the room some temperature. The hermit crab, too, could be sitting before an image of food that leads it to optimize a trajectory of its pincer, only to clunk into a TV screen. And, I suggest, we humans could be deceived by an evil demon, maximizing internal coherence while holding mostly false beliefs. In all these cases, the internal function can work properly and still bring about effects that diverge from the purpose for which that internal function was designed.

[5] I borrow the example loosely from Churchland (1995) pp93–95.

1.2.2 Epistemic norms and guidance

Now let me explain how I think this notion of an internal function can clarify issues in normative epistemology. The job of the normative epistemologist, as I see it, is to provide standards of evaluation for epistemic states (like thoughts and thinking processes). What is important about this job? One crucial answer is that such norms can give us direction for improving our thinking. That is, a key reason to determine what makes a better or worse epistemic state is that we would then like to *attain* the better ones. Put more simply, we want to be better at figuring things out; we want to be more intelligent. An important task for the normative epistemologist, then, is to figure out just what is involved in getting better at figuring things out.

If I am interested in improving my own thinking, then first I will only be happy with epistemic norms that can *guide* me toward this thinking. I do not want a norm that has no effect on me. Of course a logic teacher can have a norm for thinking, and guide me toward it using arguments, grades, and such. Or it may be that I fall under the influence of an astrologer instead, who guides me toward those norms for thinking. I might *be guided* by any number of such norms. But to guide myself toward better thinking (as I would see it) I need an internal standard. This demand, I think, is crucial to all epistemic projects that we would intuitively call *internal*:

the guidance intuition internal epistemic norms can guide thinkers from within toward achievement of that norm.

Although intuitive, the guidance intuition is not totally helpful, because this talk of "guidance" and "from within" is, so far, unclear.

To clarify some, contrast a classic standard of evaluation for mental states such as *knowledge*. This standard endorses beliefs that are *true*, as a necessary component. But

according to another common intuition from internal epistemology, the imperative to "believe what is true" cannot guide a thinker in the way we would like.[6] Chapter IV will assess this intuition, and the possibility of truth as an internal norm toward which we are guided by, say, internally available evidence. For now, let's just run with the intuition that the truth norm cannot guide us, and see where it leads us. *Why* couldn't such a norm guide us? Again intuitively: because to be *guided* by a norm from within, I need to be able to judge how I am doing with respect to it. I cannot adjust my thinking toward the direction of achieving this norm when I have no measure of where or how my thinking diverges from it. I do not have *access* to which of my beliefs are true and which false. If I somehow did, I could weed out the false ones. But as it stands the norm to "believe what is true" is not advice I can *use*, because it is not advice that could alter my mental behavior in the right way. The norm may actually be acting on me, through reliable mechanisms I have had from birth. But in order for a norm to guide me from within, I need at least a measure of error with respect to this norm that is "accessible" to me. This, too, is a common intuition from epistemology:

the accessibility intuition error with respect to internal epistemic norms are accessible to the thinker.

Again, on its own this is intuitive but not all that helpful, because it is not clear what it means for the error to be "accessible".

The guidance intuition can help us get a grip on this, however. Put it this way: a standard skeptical hypothesis has it that the same psychological state of a thinker could have mostly true beliefs in one world, and drastically false beliefs in another world (a world with deceitful demons, say).[7] If we acknowledge this hypothesis to be coherent, then we

[6] I assume that for a belief to be true involves a relation like correspondence with what that belief represents.

[7] Here and throughout I mean these psychological states to be "narrowly" construed, preferably as func-

acknowledge that the world can differ in ways relevant to achieving the proposed epistemic standard of knowledge, without some corresponding difference in the psychological state. So the mental system does not always have an internal difference to correlate with achievement of the proposed standard. My resources for improvement are the same in both cases, though my error with respect to the standard diverges greatly.

In fact, any proposed epistemic standard that demands some contingent *relation* obtain between the thinker's psychological state and something external to it will fail the guidance intuition. For since the standard is contingently relational, all the intrinsic properties of the psychological state could stay the same while its success in obtaining the relation to the things external could change.[8] Therefore, again, error with respect to this standard cannot be cognitively accessible to the thinker. The word 'accessible' gets tossed around and debated in the literature on internal epistemology, but this first gloss is simple and innocent: for error with respect to a standard to be cognitively accessible, it must have at least *some* impact on the psychological state of the thinker.

Since contingent relational properties have been ruled out, we are left concluding that only properties intrinsic to a thinker's psychological state can be candidates for epistemic norms that follow the guidance intuition—that is, candidates for epistemic norms we are intuitively calling "internal". Difference in success at attaining those norms, at any rate, will be *guaranteed* to have some cognitive impact on the thinker. Or, put another way, there will be no difference in psychological states' success relative to such norms without some difference in the mental states themselves. Or put still another way, we arrive at this other common intuition of internal epistemology:

---

tional states determined only by proximate inputs and outputs. Also I will use, following common but somewhat strange usage, 'mental state' for individual states such as a belief or feeling, and 'psychological state' for the total suite of mental states in a thinker.

[8] As Jessica Wilson points out, this talk of "intrinsic" and "contingently relational" properties is best understood as heuristic, here; what is really at issue, as we will see, is whether the standard supervenes on the psychological state.

the supervenience intuition success relative to internal epistemic norms supervenes on the (narrow) psychological state of the thinker.[9]

So far, then, with an intuitive approach to the criterion that internal epistemic norms must "guide" thinking, we have explained two other common intuitions about the division between internal and external epistemic projects. This is a good sign that we are onto something telling about internal epistemology.

The supervenience intuition has independent intuitive force, I think, arising from an epistemic intuition I think still more primitive:

the demonic intuition two thinkers forming the exact same thoughts based on the exact same experiences (narrowly construed) are surely according to *some standard* doing exactly as good an epistemic job—even if one is getting its experiences from a deceitful demon, and the other from ordinary objects.

The thinker experiencing ordinary objects would presumably have much more knowledge than the other, though, so it seems a standard of actual knowledge acquisition cannot satisfy the demonic intuition.

Though the demonic intuition probably lies beneath the supervenience intuition, the latter is too weak to capture the full intuition behind the former—and the reason is telling. Suppose, for example, that a thinker did somehow manage to have psychological states guaranteed to change with the environment. Maybe, for example, that thinker magically has a tiny patch of constant visual information that would be a slightly different color in each other possible world. Then though any epistemic evaluation of this thinker would satisfy the supervenience intuition (trivially), it would not seem to be enough to guarantee that the norm guide the thinker, as required by the guidance intuition. Suppose, for example, the psychological state of a thinker deceived by a demon differed from a counterpart

[9] James Pryor has a similar supervenience formulation of internalism in Pryor (2001) p104.

in having a tawny hazel patch where the other has a dusty rose. How could that guide the deceived thinker toward the truth, for example? Intuitively it cannot; and intuitively, therefore, that difference would not be enough to ground a different evaluative stance toward them. We would be left thinking, as in the demonic intuition, that we should evaluate both thinkers the same; the difference between them seems epistemically irrelevant.[10]

1.2.3 Guidance and epistemic agency

So what kind of difference, intuitively, is epistemically "relevant", or significant enough to warrant different evaluative stances? I think the most common and natural response to the demonic intuition is that even if some thinker is doing a horrible job getting knowledge in the demon world, we must hold that thinker *blameless* for this failure. We could not reasonably expect the thinker to do any better.

Such intuitions about "blame" can incline one toward a deontological account of internal epistemology, and indeed it is probably the most prevalent approach.[11] But here we must remember that an epistemologist might approach epistemic evaluation in several ways. It could be done deontologically, by asking whether epistemic duties have been fulfilled in thinking. Or it could be done as in virtue ethics, by asking whether the thinker exhibits the right epistemic virtues. Or, it is rarely remembered these days, it could be done consequentially—by asking how the resultant epistemic states fare relative to other possible mental states on some epistemic good. Perhaps there are other alternatives. Most internal epistemologists have gone the first route, and several are leaning more toward the second. But I think there is an important reason to keep the third in view: namely, it does

[10] This "color patch" point is inspired by an argument of David Hills' (about personal identity in Hume).

[11] Alvin Goldman makes a laundry list of deontic internal epistemologies at Goldman (1999) p116 that I will not rehearse. Earl Conee and Richard Feldman, in their own effort to sever deontology from internal epistemology, say that they "suspect that deontological arguments for internalism are more the work of internalism's critics than its supporters" (Conee and Feldman (2001) p 239 n20), though they too concede that there surely are at least some.

not rely on a problematic notion of epistemic *agency*.

What is problematic about epistemic agency? First, it is often left mysterious. By a "mysterious" epistemic agent, I mean a thinker or part of a thinker that has an unexplained (and apparently unexplainable) ability to "decide" on one thought over another when presented with alternatives. An epistemic deontology that relied on a libertarian notion of agency would presumably involve such mystery, for example. The language of deontic epistemology commonly evokes the image of a homunculus inside the head, surveying the potential justifiers that lie "before" it. The homunculus then mysteriously either pays attention to these justifiers and thus thinks in a permissible and responsible way, or else neglects these justifiers in a way that makes the thinker morally liable.

If left mysterious, such homunculi would of course be anathema to any naturalistic approach to epistemology. We can be confident, for example, that work in artificial intelligence will not ultimately rely on them. So in order not to foreclose on the possibility of AI, we should see if we can do without them. More generally, to the extent that we have such naturalization as our goal, we should try to explain how decisions get made in physical terms.

Naturally it may be that an *unmysterious* epistemic agency is compatible with deontic epistemology. But there is reason to worry here too; if we were to reduce the mystery by explaining the causal processes behind a decision, then our inclination to blame may reduce with it. Deontic approaches to epistemology seem to rely on a doctrine of doxastic voluntarism, and doxastic voluntarism seems more problematic the more the "voluntary" mechanism gets reduced. Even a doxastic form of compatibilism becomes less palatable when applied to the causal mechanisms underlying the choices themselves.[12]

Finally, epistemic agency causes problems for the accessibility constraints that internal

[12]The *locus classicus* on doxastic voluntarism as an objection to deontic epistemology is Alston (1988). For a recent overview of such issues, see Steup (2001).

versions of deontic epistemology require, as Goldman (1999) deftly argues. (I will not rehearse the arguments here.) So although I have just gestured at them, it seems there are enough problems with agential approaches to internal epistemology to warrant looking elsewhere.

1.2.4 Guidance and internal functions

Indeed, Goldman allows that a rationale other than "the guidance-deontological conception" (as he calls it) might help internal epistemology. He also points out that the deontic and guidance conceptions of epistemology come apart. He considers deontic epistemology without the guidance conception, but dismisses it (on behalf of his opponent) because it would not be sufficient to motivate internal epistemology.[13] But he does *not* consider whether a guidance conception alone might lead to internalism, and forget the deontology. This of course is just the route I would like to consider.[14]

One attractive feature of epistemic agency is that it provides a clear sense of a bad choice relative to some norm, which in turn makes good intuitive sense of the norm's guidance. But the functional notion of guidance and accessibility sketched so far in section 1.2.2 does not ultimately rely on this intuitive talk of choices. It does not even need talk of an "agent" that "knows" what facts obtain, as in typical accessibility constraints.[15] Instead, as we will see, we can think of accessible guidance in terms of internal functions. We explain doing well or doing poorly relative to this norm not in terms of duty violations, but in terms of effective performance of an internal function. Clearly this version of guidance does not rely on mysterious epistemic agency, since it could apply as well to thermostats and hermit crabs as it does to us. On the reasonable hypothesis that our intel-

[13] Goldman (1999) p130 and p116, respectively.

[14] Conee and Feldman go this route as well, I later discovered, and they argue in Conee and Feldman (2001) that it is sufficient to overcome Goldman's objections.

[15] Like Goldman's "KJ"; Goldman (1999) p117.

ligence is different only in degree of sophistication from such simple functional systems, this is a good result.[16]

Remember that the idea linking the guidance intuition and the supervenience intuition was that error with respect to a standard needs to have some impact on the psychological state of a thinker in order for the standard to guide that thinker into correcting the error. This provides an alternate explanation for why the supervenience intuition was too weak. Where the deontic epistemologist might say that a mere difference in a tiny color patch is intuitively not enough to ground differences in epistemic *blame*, we can say instead that a difference in color patch does not by itself constitute an internal measure of *error*. That is because it does not have the right causal effect, we suppose, on the rest of the cognitive system. It does not provide *feedback* toward performing some function. I propose to understand epistemic guidance as a feedback mechanism. It is hard to see how different color patches could provide guidance in this feedback sense, and systems that differ only with respect to these patches are functioning equally well.

We have seen that some creatures are such that they possess mechanisms designed to adjust their behavior in one direction or another, in order to improve their performance at some or other of their functions. The thermostat and hermit crab are two such. They possess internal functions. Furthermore, some creatures—roughly, the intelligent ones—have mechanisms that can adjust their behavioral *dispositions*, by adjusting their own mental computations. These mind-altering mechanisms are themselves part of the mind, and are themselves functions. Furthermore they too are internal functions; since the "behavior" of these functions takes place entirely within the cognitive system, their effects are internal to the creature. To alter their success or failure would require altering the creature itself, rather than merely its surroundings.

[16] In particular, I suppose, on the *computational theory of mind* hypothesis.

Of course there is also external guidance, from the external feedback of external functions. After all, people may in fact be guided toward (mostly) true beliefs by the external facts and some reliable mechanisms that they possess, just as thermostats in normal circumstances are in fact guided toward keeping the room at a comfortable temperature by their reliable mechanisms. (Poorly-designed thermostats will not sell, and so get negative feedback from quality assurance or marketing.) So a functional construal of "guidance" need not exclude what we have called external functions. Indeed external epistemology has greatly benefited from functional approaches.[17] But remember the intuitions that drove us to investigate internal epistemology in the first place. Evaluating a thinker according to its performance of external functions would satisfy neither the supervenience intuition nor the demonic intuition. The thermostat could perform its external function better or worse without altering its intrinsic properties at all. Similarly, a thinker badly deceived could be gaining much less knowledge with no change in its intrinsic properties.

External guidance would not even satisfy the guidance intuition, since its motivation was from *self*-improvement and not environment-improvement. Of course internal functions may serve external functions, as we have seen. For the individual to produce some effect may help the individual-plus-external to create some other effect. But when the individual asks what it can do, to answer that some *other* entity (the individual-plus-external) should adhere to this-and-such norm is not yet an answer. Internal epistemology demands an answer at the level of what the individual should do, which is (I have argued) to demand a goal that can provide internal feedback.

[17] I would understand "reliable mechanism" accounts like Goldman (1986) to be functional approaches; so of course are more explicitly functional external accounts like Plantinga (1993).

1.2.5 A concrete example

Here is an example to illustrate how such a standard of correctness for thought might actually be implemented in some machine. Paul Thagard, who is also interested in the area intersecting philosophy and artificial intelligence, has developed a computational model of explanatory coherence called ECHO. It takes inputs of percepts ("data"), explanatory relations, analogical relations, and contradictory propositions, and then uses these to calculate the most coherent explanation of the data available.[18] The result is an assignment of a degree of belief (or disbelief) to each of the propositions being considered. This is impressive work, but I think the most fascinating aspect is the list of preset parameters that ECHO uses to calculate its conclusions.

For example, the model has a preset "simplicity impact" parameter. ECHO diminishes the explanatory coherence of conclusions in proportion to the number of propositions required for the explanation, but leaves it up to this parameter *how much* to reduce the weights for greater complexity. There is a similar "analogy impact" parameter. ECHO also has a preset "skepticism" parameter that automatically decays each degree of belief by a certain small amount each cycle, a "data excitation" parameter for the default degree of belief assigned to the percepts, and a "tolerance" parameter that essentially adjusts the system's aversion to contradictions against its willingness to consider competing hypotheses. (At the extreme, a high-tolerance system can endorse contradictory hypotheses. Conversely, in a low-tolerance system, a hypothesis with even a slight edge in degree of belief will quickly cause disbelief in any competing hypothesis.)

The *correct* value for each of these parameters is just the kind of thing tremendously interesting to epistemologists, of course. How much should we weigh simplicity in our theory preference—a great deal, or maybe none at all? If the theory exhibits analogy to

[18] Thagard (1992) chapter 4.

other theories, is this a positive point in its favor, or a happy coincidence? How skeptical should we be about our conclusions—more like Sextus Empiricus, or more like G. E. Moore? Do our percepts warrant default, foundational justification? If so, how much? Just how bad is it to lend some credence to two contradictory propositions at once? ECHO cannot itself give answers to these questions. Thagard makes a provocative comment on the topic, though:

> In a full simulation of a scientist's cognitive processes, we could imagine better values being *learned*. Many connectionist models do not take weights as given, but instead adjust them as the result of experience ... it should be possible to extend the program to allow training from examples.[19]

That is, we would like ECHO (or, more properly, a system that incorporates ECHO) to adjust these parameters itself, learning to get *better* at its coherence calculations. The problem is, what is meant by "better"? By what standard will we compare the results of one array of parameter settings against another?

One obvious answer is: adjust the parameters until ECHO endorses the *true* hypotheses. As programmers, we have an idea of what ECHO's conclusions should be given its inputs. We can tweak and fine-tune the parameters until its conclusions match ours. In the end, we might suspect ECHO roughly reflects our own parameter settings, and this could give some insight into our cognition. But of course finding these settings for ECHO would not settle any major questions of philosophy, for it is a presumption of normative epistemology that our own settings for such parameters may not be *right*. Put another way, we take ourselves to have the ability to learn better epistemic "parameter settings". And we would like to be able to model this ability in a machine that would incorporate ECHO. This is simply the guidance intuition.

[19] Thagard (1992) p81.

How could ECHO learn its own parameter settings? We could build on top of it a mechanism that could read its output, compare it to an optimal state, and adjust its parameter settings accordingly. Of course that requires specifying what the optimal state is. In other words, a system incorporating ECHO would need an internally accessible standard for correcting its parameters. Only with such a background goal in mind could it learn on its own to adjust them one way or the other. Similarly, the presumption that we can improve our thinking via internal reflection involves a presumption that we can adjust our thinking toward some internally accessible goal state. I hope this example has made such a claim plausible; now I will try to argue for a more precise version, through the notion of learning.

1.3 Internal epistemology and learning

Only some creatures possess the kinds of mechanisms required to adjust their own cognitive dispositions. These are, I will claim, the creatures capable of *learning*. And this capacity to learn, I claim, is necessary for a thing to possess any interesting degree of *intelligence*.

Let's take the last claim first: intelligence requires a capacity to learn. Think again of the thermostat. Not to put too fine a point on it, that is intuitively a pretty dumb creature. As Dennett points out for thermostats, they do not have a notion of temperature, or the heater, or anything of the kind. They just keep a thermometer and a switch in sync, as they were designed to do. They are oblivious to whether they are measuring room temperature or outdoor temperature, and oblivious to whether they are activating a heater or a coffee maker. They are unintelligent creatures, like algae or bacteria. But imagine giving a thermostat, as Dennett suggests, further ways to detect temperature in a room, and further ways it can bring about a change in that temperature—say that it also measures air density,

and it can open and close windows. We start to think of it as "smarter", then. It does not rely so much on its external environment being a certain specific way (the boiler working, the internal thermometer reflecting actual room temperature) for success at achieving its basic function. It is more *adaptable*. And it is that greater adaptability, I think, that gives it a greater degree intelligence than its simpler cousin.

Here I take it for granted that intelligence is a matter of degree. Dolphins and chimps are smarter than octopuses, and those are smarter than squirrels, and those are smarter than flatworms, and those (it turns out) are smarter than plants. I suggest that roughly the degree of adaptability in fulfilling creaturely aims is what explains the degree of intelligence we attribute. Naturally the fancy thermostat is still at the bottom of this spectrum; maybe it is at about the level of a very simple multicellular organism, or a decent plant. To make something truly smart—like the exalted flatworm—requires that the creature have a sub-function specifically designed to take cues from the local environment in order to *improve* its performance of some other aim. More specifically, the creature should be able to *learn* how to perform its aim better given that environment.[20] Creatures that learn are creatures that meet our intuitive standard of adaptability for intelligence.

And to possess the ability to learn, I claim, is to have a cognitive mechanism that provides feedback to other cognitive mechanisms, altering them toward the achievement of the creature's aims. Creatures that can learn, in other words, are creatures that have internal epistemic norms (as they are construed in section 1.2.1). Take the case of a mouse learning to run a maze. The mouse has an aim to eat—to get the food (the quicker the better!). We have found that a mouse can improve its performance at this aim over time. And importantly, this improvement is not accidental. We can tell a causal story about how its relative ability to achieve its aim causes it to revise its performance in the next attempt;

[20]So it is no coincidence that John Pollock's recipe for building a person (in say Pollock and Cruz (1999) p179) involves adding machinery for learning—"inductive mechanisms"—at the very first step.

attempts that result in distance from the goal are inhibited, and attempts that result in proximity are encouraged. Something internal to the mouse is adjusting its probability of turning left at this corner, right at the next, and so on. The mouse apparently is able to change its dispositions to react to similar circumstances. There is some kind of feedback mechanism shaping its behavior.

Again we might be tempted to think of the cheese reward as the feedback, and thus look at the learning as a matter of external functions. Other classic cases of operant conditioning spring to mind, where the feedback takes the form of shocks, or morphine, or what have you. But, naturally, for the environment to affect a creature's behavior is not sufficient for learning. The environment affects creatures like plants, too—creatures incapable of learning. Where the sun is affects how the sunflower positions itself, but we do not think that is a case of *learning*. The sunflower has one standing disposition to react to the position of the sun. Of course the environment does provide feedback to unintelligent biological creatures through natural selection. The genes of the creature, perhaps, improve their function according to this feedback by creating creatures more capable of passing them on. But no one token plant or genome improves its functioning at some task, and thus none *learns* in this way. That is why they are not so adaptable.[21] And we can well imagine a creature very like our mouse but that never learns better ways to get the cheese, despite identical environmental feedback. The mouse does indeed improve at this external function, but only thanks to an internal one shaping its dispositions.

So for a creature to *learn* a task, it must change its own dispositions to behave (construed broadly to include cognitive behavior) in the direction of improvement at the task. The internal measure of error required for this process amounts to an attempt to assess the

[21]Compare Dennett's "tower of generate and test" in *e.g.* Dennett (1994). Plants and the like are merely what he calls "Darwinian" creatures, while those capable of some learning count as "Skinnerian" creatures and above.

difference between the creature's current state and some state the creature would "like" to be in—a state the creature has a function to reach. That is, for the creature to have a mechanism to change its own dispositions for the better fulfillment of its aims requires the creature to be built to evaluate its own success relative to a goal, and to use its relative success as feedback for altering its dispositions. The evaluation of success relative to a goal, finally, consists of two things: an implicit representation of the goal towards which the creature is trying to adjust, and an implicit representation of the current level of success with respect to that goal. Put another way, creatures that learn have at least primitive *conations* and *cognitions*. The former are, roughly, representations of how the creature wants things to be (like desires and wishes in humans); the latter are roughly representations of how the creature takes things to be (like beliefs and suppositions in humans). Creatures that learn have representations of both the world-mind and mind-world directions of fit.[22]

In fact the ability to learn may explain where representations start to split off into these two directions of fit. Ruth Garrett Millikan points out that

> Simple animal signals are invariably both indicative and imperative. Even the dance of the honey bee, which is certainly no simple signal, is both indicative and imperative. It tells the worker bees where the nectar is; equally, it tells them where to go. The step from these primitive representations to human beliefs is an enormous one, for it involves the separation of indicative from imperative functions of the representational system.[23]

The gap between bee dances and human beliefs certainly is large, but there are intermediate steps along the way. On my account, animals that can learn to perform their tasks better

[22]For discussions of direction of fit see Humberstone (1992) or Searle (1983); I base my thoughts largely on Velleman (2000). Incidentally I occasionally use cognates of 'cognition' to refer to the thinking process as a whole, rather than just the non-conative—as for example in "cognitive system". I hope this causes no confusion.

[23]Millikan (1989) p99.

can have non-propositional conations and cognitions in virtue of their internal feedback mechanisms. Learning is standardly conceived as "the acquisition of some true belief or skill through experience."[24] But to explain learning in terms of gaining belief gets things backwards, I think, since clearly there are creatures with only proto-beliefs (non-propositional cognitions) that are capable of genuine learning. I think instead the notion of learning can help explain the notion of primitive cognitions, which when propositional content is added can help explain the notion of beliefs.

But surely, you might object, the mouse did not have even an implicit goal to run the maze when it began to learn, and yet it learned to run the maze just fine. Quite right—but at least at first, running the maze is not what the mouse was learning to do. The mouse was learning to get food, and this goal it does internally represent. It also represents how it is doing with respect to that goal. When it has food, it reinforces the behavior that led to the achievement of this goal in chemically recognizable ways. Or more accurately, when it *thinks* it has food, it reinforces what led to the *apparent* achievement of this goal. The mouse's behavior would presumably change just the same way if we could somehow directly stimulate its "I've-eaten" brain centers when it arrives at the target spot. Later the mouse may actually have an implicit goal to run the maze; it has learned this new capacity or function in the service of one of its basic aims. And we would be surprised if, without any food or shocks, the mouse happened to learn just the path through the maze we wished for it. We would put this down to accident, not learning, since it is hard to see how it could be in any sense "deliberate" on the part of the mouse. To count as learning the new ability must serve some previous, internally representable aim of the creature. After all, it is not coincidence that operant conditioning always involves rewards and punishments like food and shocks.

[24] Gallistel and Glymour (1998).

The upshot is this: creatures are things that we can reasonably say have their own goals, and we rank creatures on a scale of intelligence according to how adaptable they are in achieving their goals. An especially useful tool in adapting, and so a sign of intelligence, is an ability to learn. As construed here, such learning is always done in the service of some internally representable goal of the creature's; or rather, the ability to learn in part constitutes having an internally representable goal. To learn new behavior φ is to gain the function of performing that behavior through internal feedback that provides guidance toward the performance of some higher aim or sub-aim ψ of the creature's. To learn to φ is also therefore to learn a new way to ψ. The mouse learns a new way to get food, for example, by learning to run the maze. (And this new way to get food is in turn a new way to survive, and a new way to pass on genes.)

So when we *humans* want in particular to learn to *think* better, we want to learn new cognitive behavior, and to gain new or different cognitive functions. This learning too must be guided by internal feedback; it must be guided by a background and internally accessible goal state. Such a goal state must be either another sub-aim (that perhaps itself can be improved through learning according to the feedback of an aim higher up) or else a basic aim (that cannot be so modified). In humans these basic aims might properly be called "intrinsic ends". For example, perhaps it is a basic aim of ours to get food, or avoid pain.[25] And the point for normative epistemology is that learning to think better, like learning to do anything, requires improving according to the internal feedback of matches between cognitions and conations. In other words, the ultimate *internally available* standard for better thinking—the standard that would permit us to *learn* better thinking—is the better (apparent) satisfaction of our ends.

[25] As I will argue in chapter II, though, any one such aim can be overruled by sufficiently many others, according to a comprehensive calculation of coherence.

1.3.1 Learning and intelligence

Let's look briefly at a series of fanciful examples. On the one hand, the examples should demonstrate that our natural inclinations for evaluating thinkers incline us to prefer the ones that learn in the sense outlined. On the other hand, the examples will help illustrate an objection to my construal of learning: namely, that if we construe learning as a process of induction instead, then learning does not require background aims. I disagree, and will argue in section 1.3.2 that the inductive process too depends on a creature's internally representable aims.

But first, the examples. Consider two brains in vats, Amanda and Bill. They were both captured by aliens shortly after birth. In the vats their afferent electrical and chemical impulses are adjusted magnificently by the clever aliens to simulate for them a life very like the one they would have led had they not been captured.

Amanda does what we would ordinarily consider a good epistemic job. Of course her earnest cognitive work is not getting her many *truths*—but we are likely to think that, at least in some sense, she is no worse a thinker for that. Bill, meanwhile, is a pathological wishful thinker. Whenever he desires that p, he comes to believe that p—or at least he attempts to form and sustain that belief as far as possible. Bill is doing no worse than Amanda when it comes to getting truths, we can suppose, but we are still likely to think that he is the worse thinker of the two.[26]

Why? What is so wrong with wishful thinking that we are still tempted to fault Bill, even though presumably *no* epistemic strategy will get him (empirical) truths? Is wishful thinking simply an irreducible epistemic sin? I do not think so. For consider two more brains in vats, Carl and Denise. These two were lucky enough to be captured (at birth) by

[26]The example is reminiscent of one in Pryor (2001), though he points out that internal epistemologists have appealed to such brain-in-a-vat examples since at least Cohen (1984).

relatively beneficent aliens, who attempt to make up for their atrocity by replicating for their abductees what they think must be an ideally pleasant environment. The aliens go about this by determining Carl's and Denise's desires, and then creating for them the right inputs to their brains for them to believe that their desires are satisfied. The aliens are very good at this.

As a result Carl has developed a wishful thinking mechanism, forming the belief that p whenever he desires that p. Can we blame him? It seems he has good reason to think that he lives in a world where all his desires are satisfied. Every time he has wanted something to be the case, he has soon come to believe it is the case. He hopes there is a ferris wheel behind him, and when he turns around, sure enough he has such an experience. He hopes there will be peace on earth, and comes to read in the paper that there is. And so for each of his desires, until eventually he does not have to turn around or read the paper—he just makes a reasonable assumption. He also wants there to be undetectable pixies, and comes to believe in them too. He does not give himself explicit inductive reasons to believe in the pixies, though; instead, it is simply a result of his wishful thinking mechanism. Denise, on the other hand, will not form the corresponding belief when she notices a desire. She waits until she has the required experience—which she inevitably has, except in cases like *her* desire for undetectable pixies. She does not have experience of those, of course, and so she does not believe in them.

Which brain is smarter? I think Carl's. It has adapted better to a (simulated) environment radically different from ours. Both Carl and Denise have good (subjective) reason to believe there are undetectable pixies, and only Carl is savvy to it. Put it this way: suppose instead of being brains in vats, they were both actually in a world where desires that p thereby brought it about that p, just as a law of nature. Denise would be ignorant of this law, while Carl would have learned it. But we can suppose that world is experientially

identical to what they get in the vats, and so they should come to think the same way. In either situation Denise should notice the potential success of a wishful thinking mechanism, just as Bill should notice its failure.

Just what is meant by "success" and "failure" here? Well of course that is the question. It clearly does not have to do with procuring truth. I have already suggested it has to do instead with how well their thinking manages to serve their further goals—or rather, how well their internal learning functions manage on the standard of matching the cognitions to the conations.[27] This explains both why Amanda is a better thinker than Bill, and why Carl is a better thinker than Denise. It also explains why wishful thinking is wrong both in Bill's simulated world, and in our own. Wishful thinking is unintelligent if a better adaptation to the (apparent) local environment would do without it, where a "better adaptation" is understood in terms of apparent instrumental success.

1.3.2 Learning and induction

But an alternative explanation for these brain-in-vat intuitions is to say, for example, that Amanda and Carl are simply better at predicting experience. They have formed inductive inferences (implicitly or explicitly) that Bill and Denise have not.[28] Amanda predicts her experiences reasonably well, while Bill predicts his favorite curry for dinner and often gets pasta instead. Carl predicts there will be a ferris wheel behind him, while Denise fails to. Understanding learning this way would also account for the greater degree of adaptability, and thus greater felt intelligence in Amanda and Carl. And their skill at predicting

[27] Wait a minute! Isn't Denise doing *exactly* as well as Carl when it comes to this standard? No. For example, Denise does not get to believe in pixies. Closer to home, she wants there to be money in her bank account, but wastes time actually checking to see—time she could have spent better. Denise also wants people about whom she has no information to be happy, but does not believe they are. And so on.

[28] In the background I am picturing *formal* versions of inductive learning—such as updating degrees of belief according to Bayesian rules, or a "logic of inquiry" approach like that explored in Kelly (1996). These approaches are of interest to me since they have notable potential for artificial intelligence. But I think my points would apply equally to other versions of inductive learning.

experience is a measure independent of background goals, we might think. Either they are correctly predicting experience or they are not, whether or not they want to. And, we might think, they are smarter to the extent they can catch on to the environment around them, and predict it—*whatever* their goals are.

On this version of the inductive learning proposal, predicting experience is a primitive epistemic good. But consider two more brains in vats (why not). Erin's and Fred's brains are the victims of a renegade, demonic alien who likes toying with them. In particular the alien has it rigged so that whenever it detects a prediction of experience on the brains' parts, it provides an experience to contradict it. As a result Erin's experience-prediction mechanisms have long ago withered away as merely frustrating. She has found other ways to cope, such as "predicting" experiences that she does not want. Meanwhile Fred continues to predict his experiences, and gets them wrong every time. Again, one of these brains is failing to learn, and I think that brain—Fred's—is the less intelligent one. Of course this lack of intelligence may simply be from hardwiring; perhaps Fred cannot help but attempt to predict his experiences, because he is just built that way. He is built, that is, not to be able to adapt to a circumstance where predictions do not serve him.

Of course we can construe Fred's failure to learn as another failure to induct, this time a failure to induct from his failure to predict. And I agree that learning can often, or perhaps *always*, usefully be understood as an inductive process. What I do not agree with is that through induction we can understand learning independently of a creature's aims. Put it this way: what is Erin *correctly* inducting? In what sense did Erin learn that her experience-predictors *fail*? To learn that her prediction mechanism fails her, she would need a mechanism overlooking it and adjusting it according to its performance. Suppose what is monitoring the experience-predictor spits out a signal like 1 when an experience matches a prediction, and spits out 0 for a mismatch. Why should Erin's brain be set to

encourage the experience predictor when it reads a lot of 1's, and inhibit it upon reading 0's? Why not the other way around? Presumably this is because of Erin's ends. She tries to predict for a reason; prediction is meant to serve higher aims, and when it does not she is adaptable enough to learn not to use it. If some creature in our world adjusted such a mechanism toward getting *more* mismatches, we intuitively would not call this learning or intelligence, even though it is induction. That is because in this world matching predictions serve aims, and mismatching ones do not.

Put in terms of induction, there have to be at some point preset categories that Erin simply, primitively inducts upon—maybe ones like "pain-producing" or "food-obtaining". The categories on which she is primitively designed to induct set the standards of success and failure for her derivative inductive mechanisms. Another way to look at it is through the impetus to inquiry. Suppose we can correctly describe all learning as induction, and suppose also that creatures *do not* start biased with a small number of primitive, preset categories. Instead suppose there is a huge multitude of categories for inducting, such as "stimulation type 39 of optic nerve 781", and suppose the induction starts with an indifferent distribution of probabilities for these categories. Then the creature's probability distribution will have *huge* entropy (as they call it)—that creature will be completely uncertain about everything, and has no way to pick one hypothesis over another. And the question then becomes: what drives the creature to change those probabilities, investigating some hypotheses rather than others, in an effort to reduce this entropy? Presumably it will be built to inquire only into certain matters of interest to it.

In summary, then: intuitions from brain-in-a-vat examples lend credence to the idea that an ability to learn by adjusting to the apparent environment is a mark of better thinking. And interpreting this learning as an inductive process does not excuse us from the conclusion that the internal standard of learning is a matching of cognitions to conations.

1.4 The normative status of the standard

The argument so far takes this skeletal form: only creatures—things in possession of the right kind of functional makeup—can be intelligent. Our intuitive attributions of intelligence track the adaptability of the creatures to their environment, and the capacity to learn is especially powerful in providing such adaptability. The capacity to learn, furthermore, is the possession of an internal function to adjust behavioral dispositions (including cognitive ones) through feedback in the direction of the creature's aims. This internal feedback gives the creature, implicitly, primitive conations and cognitions—representations of how the creature wants the world to be, and how the creature thinks the world is. In particular creatures with the capacity to learn to think better (like humans) must do so according to some internal feedback mechanism. The resulting picture satisfies major intuitions of internal epistemology without obviously relying on any mysterious version of epistemic agency.

So the proposed internal standard of correctness for thought is that of apparent goal satisfaction. According to this proposal, internal epistemology is *pragmatic* epistemology.[29] In the next chapters, we will see how external constraints on goal satisfaction, both from the conative and cognitive side, restrict the kinds of thinking that would best bring it about. (For example, wishful thinking is not an effective way to think, even by this standard.) I will also show how such a pragmatic epistemology can do without a naive, instrumental version of practical reason. The result will be a "foundherentist" approach to mental computation. The approach has a natural algorithm to make it more precise,

[29]Crass and shocking, I know. But strangely, many epistemologists who wish to connect up with artificial intelligence are pragmatists. Thagard for example defends a pragmatic approach to scientific realism in Thagard (1988). Patricia and Paul Churchland are pragmatists. And hidden away in Pollock and Cruz (1999) are claims like "the ultimate objective [when evaluating cognitive architectures] is not truth, but practical success" (p175). There could be a merely *sociological* explanation for this tendency, though I think I have given a good philosophical one.

and that algorithm has a natural architecture to instantiate it. The result will help patch normative epistemology into cognitive science, which I think is all that can be asked of "naturalized" epistemology.

Meanwhile, what is the normative status of this proposal? For one thing, I am attempting to describe in more detail the thinking humans do in virtue of being intelligent. That involves describing the standard according to which I think humans must, ultimately, adjust even their thought-forming mechanisms. If they *must* so adjust, where is the normativity? In a system's relative success or failure at achieving this standard. Such failure or success, for example, can explain our intuitive attributions of rationality and irrationality in the brain-in-vat cases—intuitions fundamental to internal epistemologists. In attempting to describe intelligence, I am engaged in a normative endeavor, since calling something intelligent is at least partially evaluative. For another thing—or, for another way to view the same thing—I am making a *recommendation* about how to go about building smart robots.

Finally, I am indirectly endorsing a philosophical program, advocating a more hearty acceptance of epistemic pragmatism (properly understood). I have claimed essentially that given the kind of creatures we are, and given what it means for such creatures to learn, we cannot *help* but be epistemic pragmatists. This is no horrible thing; rather, it should be embraced, for it means the furthering of all our *rational* goals. And philosophical problems that ignore pragmatic implications—or look for a standard beyond them—should be reexamined for their point.

CHAPTER II

WISHFUL THINKING AND COHERENCE EPISTEMOLOGY

So far I have argued for an internal, pragmatic epistemology that at least has potential for being naturalized. In this chapter I argue that achieving such an internal standard must involve achieving a *coherence* among all thoughts, both cognitive and conative. First, though, I look at a completely natural objection to make at this point against any internal pragmatic epistemology: namely, that it seems to endorse wishful thinking. My response to this objection leads smoothly to the topic of coherence—and thus to a more specific form of internal pragmatic epistemology that will bring us still closer to a computational model.

2.1 Wishful thinking

A pragmatic epistemology holds that thoughts are better according to whether they contribute to desire satisfaction for the thinker. (Contrast, incidentally, a pragmatic *meta-physics*, which holds that *truth*—rather than something like justification—depends in some way on instrumental success. I will often speak loosely of "pragmatism" when I mean the epistemological variant.) An *internal* pragmatic epistemology (like mine) construes desire satisfaction internally, so that a thinker's desire is "satisfied" when the thinker believes

that the desire has been attained, whether or not it actually has been. Thoughts are better when they contribute to apparent desire satisfaction, or what we might call "subjective happiness".

But this just sounds like a solipsistic and tender-minded incitement to epistemic irresponsibility, amounting to advice along the lines of "believe what you want, and evidence be damned." Since I have a desire to be an adventurous pirate, any beliefs I might form to the effect that I *am* one appear by internal pragmatism to be justified. In other words, the position seems to countenance and even encourage that worst of fallacies, wishful thinking. But any normative epistemology that endorses such an obvious fallacy hardly deserves the name. Whims would count as reasons; if you want there to be pixies, a belief in pixies is thereby justified, and if the next moment you wish there were elves instead, a belief in elves is now best. If two people or cultures have contradictory desires, then the contradictory beliefs are both justified; for that matter if within one person can exist both a desire that p and that $\sim p$, then the associated contradictory beliefs are both justified.[1] Pragmatism of this sort seems to entail that "anything goes" when it comes to evaluating thoughts; it seems, in effect, to give up on normative epistemology altogether.

2.1.1 Epistemic relativism and "anything goes" epistemology

In fact wishful thinking seems to come part and parcel with a general permissive relativism that gets associated with pragmatism. But it is easy to conflate the following three objections to an epistemology:

1. It condones wishful thinking.

2. It is relativist.

[1] Here and throughout I am using 'justified' in what is intended to be an ecumenical way, as a kind of shorthand for "good according to some epistemic norm."

3. It is such that "anything goes".

Stephen Stich is sensitive to similar objections in his argument for an external pragmatic epistemology.[2] Though wishful thinking is not such a concern for his position, he carefully separates the other two points. Then he grants that his position leaves epistemic evaluation highly relative to circumstances and goals, and does not consider it a problem for his view. He points out that even truth-based, consequentialist approaches will involve similar relativism.[3] But then, according to the "anything goes" objection, for pragmatism to commit to relativism "simply gives up on the project of distinguishing good cognition from bad."[4] Wishful thinking gets implicated because it appears to be the inevitable result of such epistemic promiscuity—no belief is intrinsically better than any other, so when faced with the task of choosing one, believe what you want. Stich's response:

> Pragmatism does not give up on the project of assessing cognitive processes. Quite the contrary. Epistemic pragmatism offers an account of cognitive evaluation that is both demanding and designed to produce assessments that people will care about. ... It will near enough never be the case that pragmatism ranks all contenders on par.[5]

This passage applies to internal pragmatists, too—for only some beliefs will lead to *apparent* desire satisfaction! At least, then, I have reasons not to believe *some* things, such as things that I do not want to believe. Therefore not *anything* goes.

Still, this hardly rules out wishful thinking. The epistemic pragmatist (Stich included) *does* claim that no belief is intrinsically better than any other. He does not conclude

[2] Stich (1990).

[3] For one thing, what cognitive processes garner truth will be relative to whether or not you are in, for example, a perverse demon world. For a more grounded example, it seems that whether to be an intellectual maverick or a conservative in your academic field might well depend, in attempting to maximize truth, on the current relative percentage of mavericks and conservatives in your community. See Stich (1990) pp136–140.

[4] Stich (1990) p141.

[5] Stich (1990) p141.

that you should therefore believe what you want, but he does conclude that you should therefore believe what "would be most likely to achieve those things that are intrinsically valued," usually by the believer.[6] When the internal pragmatist replaces 'achieve' with 'apparently achieve', wishful thinking seems to follow.

2.1.2 The wishful thinking objection

It is not hard to see how. Compare two belief-forming mechanisms: one forms beliefs according to ordinary practices of evidence evaluation and the like, and the other always wishfully thinks. That is, it automatically forms the belief that p for any desire that p, and gives the belief weight that trumps any other belief formed by any other mechanism should there be competition. We are to evaluate these systems according to a standard of apparent desire satisfaction, which is roughly to evaluate the ratio of matches between desires that p and beliefs that p. It seems clear that the wishful thinker would have a serious advantage in this evaluation.[7]

Naturally such a pure wishful thinker would never live very long—at least, a human one wouldn't. When hungry he will simply form the belief that he has eaten rather than get food, and when thirsty he will simply form the belief that his thirst has been quenched. He would never move, and be dead within days (faster if there are tigers). This is little defense against wishful thinking for internal pragmatism, however. First, the wishful thinker may not desire a long life, and our evaluation of his cognitive system can only be done relative to such desires (and perhaps we shouldn't blame him, if he can truly live his short days in such ecstasy!). Even if he did desire a long life, he need merely *believe* that he will enjoy one, according to an internal pragmatic standard. Luckily for him, if he does have

[6] Stich (1990) p131.

[7] Contrast Bill, the wishful thinker in a vat from chapter I—we suppose in that case that at some point he is not *able* to maintain his fanciful beliefs when faced with blatant evidence to the contrary. He is not a good learner because he does not notice that his beliefs are always being thwarted in this way.

the desire to live a long life, he also believes that he is living one. He thus is remarkably effective in gaining apparent desire satisfaction, and so by internal pragmatic standards his cognitive system is top-notch. True, he may not be able to contain all this epistemic goodness for very long. That is, he may not be able to garner very much of it during his short lifetime. But first, it is not clear that some *quantity* of desire-satisfaction is the appropriate goal to judge cognitive systems by; a cognitive system that brought about tiny bits of desire satisfaction every year or so over the period of eons is not obviously preferable to a high concentration of less total subjective happiness over a shorter amount of time. And anyway, perhaps a somewhat less pure wishful thinker would last longer at the expense of mere 98% desire satisfaction, say—and perhaps somewhere on that spectrum is a maximum of subjective happiness. Still, even if *pure* wishful thinking is not the ideal strategy for subjective happiness, it looks like wishful thinking will be a heavy factor in the race for the pragmatic trophy.[8]

2.1.3 Wishful thinking as rational

One possible response for the internal pragmatist is to include the wishful thinking fallacy with the other purported "fallacies" that epistemic pragmatists such as Stich and Gilbert Harman claim are not, in the big pragmatist picture, irrational practices after all. Take for example the documented phenomenon of "belief perseverance", where undermining the evidence that led to a belief fails to undermine the subject's conviction in that belief. This strikes us as an irrational thinking strategy that could land one in real, pragmatic trouble. But this tendency may simply be an unfortunate side effect of a more general cognitive strategy that turns out to be on balance pragmatically *wise*, given limitations like the size of our brains and the time required to remember things: namely, the strategy to

[8]Here, in the wishful thinking section, is a good place for a disclaimer: although many pragmatists use the view to defend religious faith, I have nothing like that up my sleeve. In fact I am an atheist.

retain a conclusion and discard the reasoning behind it as (typically) a waste of effort. We can all think of cases where we recall the conclusion of a paper or editorial or theorem without recalling offhand how that conclusion was reached. Instead we keep a kind of placeholder: "there were reasons for this belief I found convincing at the time." And we might on reflection endorse such economization. The belief perseverance phenomenon is a foreseeable negative side effect of this strategy—a negative side effect that simply gets outweighed by its pragmatic benefit. Thus it turns out in the grander scheme of things belief perseverance can be seen as part of a rational belief strategy.[9]

A paper by Dion Scott-Kakures might help the internal pragmatist in this way with wishful thinking.[10] Scott-Kakures draws on contemporary psychological studies to form a general account of "motivated believing"—a category meant to include both wishful thinking of the type we are examining, and "unwelcome" believing such as hypochondria, where feared things are believed. A unified and psychologically respectable account of these, Scott-Kakures claims, is one

> according to which motivated believing is in the service of the realization of the subject's goals and values, and according to which there is no bright line between motivated and (what we're apt to regard as) unmotivated or accuracy-driven cognition.[11]

The putative justification of wishful thinking, as in the case of belief perseverance, is fairly simple. We have to make evaluations and decisions under conditions of uncertainty and limited resources, using probabilistic reasoning. A standard and perfectly reasonable-seeming practice in such evaluation is to set confidence thresholds that need to be exceeded

[9] See Stich (1990) pp152–3, which relies in turn on Harman (1986) pp37–42. Here and throughout I use 'rational' ecumenically, to mean "good thinking according to some epistemic standard."

[10] Scott-Kakures (2000).

[11] Scott-Kakures (2000) p350.

before an evaluation can be considered complete. Now let us take on board, as the pragmatist would, a "pragmatic hypothesis testing account", according to which

> how intensively and in what manner a subject tests a hypothesis will reflect her values and interests. Cognition, on this view, is suited for the securing of rewards or benefits and the avoidance of what is undesired.[12]

As a descriptive claim about rational agents, it is surely uncontroversial that for example we will spend more time evaluating the truth of beliefs that are genuinely important to us than, say, truths about Sebastian Cabot. And it seems rational we might do so. But from this principle of rational belief evaluation—take costs and benefits of the investigation into account—seems to follow methods that are not different in kind from extreme cases of wishful thinking. Some clear costs to the investigation are the time and effort spent. But crucially, also among the costs of the investigation are costs for error—getting a false positive or false negative.

> Of course, [the cost of error] is not surprising, since most of us are apt to agree that believing truths typically confers some benefit. But what is distinctive about such accounts is the claim that, with respect to many questions, the cost of false positives, on the one hand, and the cost of false negatives, on the other, will not be identical; such error costs are asymmetric. The costliest error, the "primary" error, is the error the subject is preponderantly motivated to avoid. This error is fixed by the aims, values, and interests of the cognizer. Thus, in so far as a subject's hypothesis testing is sensitive to the avoidance of the costliest error, it serves to bring about what that subject values.[13]

[12] Scott-Kakures (2000) p362. In a footnote to this passage Scott-Kakures quotes Stich: "the pragmatic account urges that we assess cognitive systems by the likelihood of leading to what their users value" (Stich (1990) p136).

[13] Scott-Kakures (2000) p363. He cites some psychological studies to back up the claim about asymmetric error costs.

The putative defense of wishful thinking as rational, then, goes like this. Since error costs are asymmetric, it is thus (pragmatically) rational to place confidence thresholds for coming to believe p or $\sim p$ asymmetrically. Someone dying of thirst *should* have a low confidence threshold for believing the drink offered him is not poisonous; a false positive for the "it's poisoned" hypothesis would be too costly. The queen with a treacherous court, on the other hand, has worse costs associated with a false negative for the hypothesis, and thus should rationally require a lower confidence threshold for believing her drink is poisoned.

In the case of wishful thinking: suppose you desire that p. To see if your desire is satisfied, you evaluate the hypothesis that p. But like all hypotheses, you undertake this evaluation probabilistically and under limited resources. Since you desire that p, you will be disappointed, disheartened, or otherwise saddened if you believe $\sim p$. As Scott-Kakures puts it, "there is an intrinsic cost built into the rejection of a desired hypothesis: negative affect."[14] Thus all other error costs being equal, the false negative (concluding that $\sim p$ when in fact p) has a higher cost for you than a false positive (concluding that p when in fact $\sim p$). Set your confidence thresholds accordingly, and lo and behold, you have *rationally* made it more likely you will conclude p, just as a result of desiring that p.[15]

[14] Scott-Kakures (2000) p366.

[15] It is possible to read some of this view back into William James, though I leave the real exegesis to the reader. Here are two of the most ripe quotations:

> The question next arises: Are there not somewhere forced options in our speculative questions, and can we (as men who may be interested at least as much in positively gaining truth as in merely escaping dupery) always wait with impunity till the coercive evidence shall have arrived? It seems *a priori* improbable that the truth should be so nicely adjusted to our needs and powers as that. In the great boarding house of nature, the cakes and the butter and the syrup seldom come out so even and leave the plates so clean. Indeed, we should view them with scientific suspicion if they did (James (1896) p201).

Though this next passage is aimed toward justifying religious views in particular, the general pragmatist point is again like Scott-Kakures's (the emphasis is as in the original):

> Scepticism, then, is not avoidance of option; it is option of a certain particular kind of risk. *Better risk loss of truth than chance of error*—that is your faith vetoer's exact position. ... To preach scepticism to us as a duty until "sufficient evidence" for religion be found, is tantamount therefore to telling us, when in presence of the religious hypothesis, that to yield to our

A critic might object that these cases are not truly ones of wishful thinking. In these cases the believer does test the hypothesis, even if the results are somewhat biased (and *perhaps*, rationally so). The believer is showing at least some interest in "how things really are." In genuine wishful thinking of the type that should worry pragmatists, the critic might say, a believer does not test hypotheses at all, but directly forms the beliefs that conform to his desires. A pragmatist could respond to this critic that the "genuine" wishful thinking cases differ only in degree, and not kind, from the "rational" ones. Demanded confidence thresholds can get lower and lower for the preferred result, until any degree of confidence is virtually guaranteed to pass the threshold. The cases of genuine wishful thinking may just be an extreme of what is a generally rational approach to hypothesis testing.

Such a response hardly shows wishful thinking to be *rational*, however. First of all, it is not clear that pragmatic hypothesis testing as a general method is rational. Scott-Kakures's cases include what would normally be considered paranoia ("unwanted", not wishful, believing) and unreasoned optimism—as for example the study he cites in which "the correlation between an individual's judgment of his own physical attractiveness and others' judgments of the same individual's attractiveness is .24."[16] It is hard to call such cases rational. But even if there is a good case for the rationality of pragmatic hypothesis testing, and even if genuine wishful thinking is different only in degree from acceptable cases, that might show only that the border of irrationality is fuzzy—not that there isn't one. Setting a margin of error so wide that the hypothesis is guaranteed to pass is surely irrational by anyone's standard.

Perhaps though, like the belief perseverance cases, genuine wishful thinking is an irrational result of a generally rational *mechanism* that has occasional unfortunate but out-

fear of its being error is wiser and better than to yield to our hope that it may be true (James (1896) p205).

[16] Scott-Kakures (2000) p348.

weighed negatives. Just as we do not consider it rational to hold onto a belief after its evidence has been undermined, but *do* endorse the mechanism that sometimes brings about such results, so we might consider it irrational to think wishfully, but endorse the mechanism that occasionally leads to it. Such an argument may be tempting, but there is an important disanalogy between the wishful thinking and belief perseverance phenomena. Incidents of genuine wishful thinking do not seem to be a necessary potential cost of adopting pragmatic hypothesis testing; rather, they look more like a hijacking for ulterior motives of an otherwise rational mechanism. Pragmatic hypothesis testing says: adjust your confidence intervals based on the cost of mistakes (and these intervals may be asymmetric if the costs are). This procedure is rational as long as *actual costs* are reflected in the calculation. It is not rational for me to have a minimally low confidence threshold for tests of whether I am a pirate, for example, no matter how much I want to be a pirate. It is rational to have the acceptable margin of error somewhat skewed in favor of the pirate hypothesis, but only by a minimal amount geared to the proper disappointment should I discover I am not one.

The internal pragmatist who wishes to defend wishful thinking as rational may have recourse to one other response. Talk of *actual* costs is too external. Perhaps cases of genuine wishful thinking appear irrational because we do not share the epistemic perspective of the believer; instead we take the perspective of an outsider who knows the truth and sees it sadly neglected by a deluded and biased belief. As Scott-Kakures says,

> from an external perspective, possessed of the truth, it is easy to survey the situation of the motivated believer and to characterize her reasons for believing p as bad reasons. But the engaged-believer has no such guarantee, and she has plenty to lose.[17]

[17] Scott-Kakures (2000) p370.

That is, we do not share the desires and epistemic pressures of the agent apparently thinking badly, and therefore cannot sympathize with her plight. But Scott-Kakures notes that the agent herself—even when she later shares our third-party perspective—might look back on such decisions and still, remembering the pressures involved, endorse the testing process she actually underwent. Perceived costs can be higher or lower than real costs, of course, and in the moment all the believer can do is act according to the perceived costs.

2.1.4 Wishful thinking as irrational again

Scott-Kakures concludes "a policy that weighs the costs of errors so as to bias the subject against the costliest mistake is far from irrational."[18] But still, there will be genuine wishful thinking cases that even Scott-Kakures feels compelled to call irrational. Suppose I believe I am a pirate because my perceived cost of not being a pirate is enormous, the end of the world; to discover I am a mere philosophy student would be a fate worse than death. No cost could be higher for me, then; if I am not a pirate all is hopeless anyway, so there is only negligible further cost to believing falsely that I am. Given my perceived asymmetric costs, it may seem Scott-Kakures is committed to saying it is rational for me to set the minimally low threshold for belief that I am a pirate. He does not want to bite this bullet, however—and who can blame him? In such cases he wants to say instead that thinkers like me are being irrational by "paying scant attention to errors about which they really do care."[19]

But notice this response is *prima facie* a bit weird: that I am paying scant attention to errors about which I actually care strongly suggests that though maybe I believe I desire being a pirate above all else, in some important sense I actually *don't* desire it as much as I believe—not enough to warrant the bizarrely skewed confidence margins. Perhaps the

[18] Scott-Kakures (2000) p370.
[19] Scott-Kakures (2000) p370.

claim here is that I have inaccurate beliefs about the costs, and these inaccurate beliefs are skewing my hypothesis testing inappropriately. Perhaps I believe (falsely) that being a pirate is the only way to make a living these days, and I desire vehemently to make some kind of living for myself. Naturally I conclude that what I most desire is to be a pirate. But in this case my desire is a simple result of mistaken means-end reasoning. I just desire a decent job of some kind, and this does not (in today's economy, from what I hear) amount to a desire to be a pirate. In this sense perhaps I do not *really* desire to be a pirate after all.

For one thing, though, this response is not easily available to the *internal* pragmatist. More importantly, the mistake in my reasoning *might* be that straightforward, but then again it might not be—it seems to me I might *really* desire very strongly to be a pirate, and my belief about this desire might not be mistaken. Scott-Kakures seems strangely confident that no one could have such desires, since he thinks *anyone* abusing pragmatic hypothesis testing in this way is "paying scant attention to errors about which they really do care." Perhaps this confidence comes from the view that no rational agent could care that strongly about such things, as a fact about rational agents; in other words, perhaps Scott-Kakures feels it is not my belief about the means to my desire, but rather my *desire* itself, that is mistaken or irrational.

This need for evaluation of desires is crucial in responding to the wishful thinking objection. Realistic desires, intuitively, are not the type that require wishful thinking in the first place. But irrational desires are not only awkward for internal pragmatism; *any* pragmatism that straightforwardly evaluates cognitive processes by their success at bringing about desire satisfaction will have these difficulties. Consider an external pragmatism like Stich's, and suppose an agent intrinsically values ignorance and error. Such an agent would have "justified" beliefs that are systematically bad by any intuitive standard.

More to the immediate point, invoking irrational desires as a response to genuine wish-

ful thinking is awkward for Scott-Kakures's account. If we can call my pirately desires irrational, we can speak similarly of anyone's perceived costs from our third-person perspective. Perhaps wishful thinking would be rational along Scott-Kakures's lines if we could always count on having appropriately rational desire structures. But meanwhile Scott-Kakures cannot have it both ways. If my thinking I am a pirate is irrational on grounds of mistaken desires, then why not say the same for more innocent cases, like those inclined to think themselves better-looking than they are? Perhaps it is also irrational to want so much to be good-looking that you are willing to bias your hypothesis testing. Perhaps in general it is irrational to want something that much. Or perhaps the rational thing to do in belief formation is always to weigh the evidence without taking your (possibly irrational) desires into account at all. This is certainly a common intuition, and we are led back to it merely from the innocent idea that some desires—such as wanting to be a pirate—are irrational.

In summary, Scott-Kakures's paper looked like it provided the best shot for defending epistemic pragmatism against wishful thinking charges. But it does not seem to work. His argument only shows that it can be marginally *more* justified to form the belief that p given such a desire. First of all, this conclusion may not be right in the specific way needed to justify some wishful thinking. Though costs for error can be asymmetric, and confidence margins rationally set appropriately, intuition suggests the particular cost of negative affect from disappointment should not bias against a false negative *at all*, let alone enough to guarantee belief. But even if Scott-Kakures is right, he only shows there are *some* cases where *some* bias from wishful thinking is permissible; in the end Scott-Kakures agrees, too, that genuine wishful thinking is irrational.

2.1.5 Thinkful wishing

If internal pragmatism leads inevitably to the conclusion that it is rational for me to believe I'm a pirate, given such a desire, then this can only reduce internal pragmatism to absurdity. But from what has been said so far, wishful thinking still appears to be a way to form justified beliefs according to the internal pragmatist.

In fact there is still worse news for the internal pragmatist. Even with an answer to the wishful thinking problem, there will be another problem in the opposite direction. For here is another cognitive method that appears to garner "justified" beliefs by pragmatist lights: suppose you have a belief that p. Then form the desire that p. For example, suppose you believe armageddon is approaching. Then simply desire that armageddon be approaching. *Voilà*, your beliefs have led to desire satisfaction, and thus are justified according to internal pragmatic epistemology. Call such an approach to cognition *thinkful wishing*.[20] Note that this problem is not unique to the internal pragmatist. Instead of matching desires to beliefs, someone could thinkfully wish in the external sense by desiring things to be just as they *are*, and *voilà*, that person's beliefs would be justified. External epistemic pragmatists like Stich would presumably need an answer to this concern just as much as the internal epistemic pragmatist does.

True, thinkful wishing is not a belief-forming mechanism (since it leaves the beliefs alone and fiddles with the desires instead), but it is a *thought*-forming mechanism that will bring justification to beliefs according to the internal pragmatist. Ruling out thinkful wishing on the grounds that pragmatism looks for good *belief*-forming mechanisms seems unfairly *ad hoc*. Beliefs are in the service of desire-satisfaction, to a pragmatist, and beliefs affecting desires in this way are apparently doing a great job at bringing about desire

[20]Thanks particularly here to Eric Lormand for discussion of thinkful wishing; we independently found interest in this symmetry, but the phrase is his.

satisfaction.

Perhaps an internal pragmatist could claim that beliefs must bring about the satisfaction of particular, antecedent desires to be justified—not just bring about any old instance of desire satisfaction. But by this understanding of apparent desire satisfaction, it seems, desires must be immutable; beliefs cannot rationally bring about change in desires. For if they did, they would presumably fail to lead to the satisfaction of the old, particular desires that were revised, and thus be unjustified.[21] If we allow that beliefs can (rationally) influence desires, and that such beliefs can be justified, then chronological priority of desires over a belief will not generally be required for the justification of that belief.

For example, I desire to be a pirate. But I have contravening beliefs such as that this is the 21st century, or that those ships are outdated and would never survive the coast guard, or that the pirate life is not in reality so glamorous after all, or that parrots are too expensive. Intuition suggests these contravening beliefs should alter my desire in some way; they should show that my desire is not so *realistic*. But a pragmatist interested in satisfying particular, antecedent desires would have to say instead that those contravening beliefs are relatively unjustified, because they frustrate forming the belief that I am a pirate. Of course a pragmatist would *want* to say, intuitively, that these beliefs are justified in part because they lead me to concentrate on desires that are more likely to be fulfilled. In other words they contribute to *general* desire satisfaction. But if general desire satisfaction is the pragmatic goal—if the beliefs as a whole are to be measured in terms of their contributions toward the desires as a whole—then thinkful wishing will be as much of a problem as wishful thinking.

[21] This argument applies to the adjusting of intrinsic desires as well as to intermediate, means-end ones—if people can reflectively change their intrinsic desires over time. The same might *not* apply to a view that involves antecedently fixed aims of the thinker. A pragmatism grounded in these may fare better against the thinkful wishing charge, and indeed my own view takes something like this route.

2.1.6 A way out: spontaneous thoughts

The symmetric structure to these problems suggests a way out. Why not believe p given the desire that p, or desire that p given the belief that p? This is only an issue when there isn't already a match between the content of what is desired and the content of what is believed. So suppose some thinker has a desire that p and a belief that $\sim p$. This is, all else being equal, an unfortunate state of affairs for a thinker—especially according to an internal pragmatist chiefly concerned with apparent desire satisfaction. To alleviate the situation requires either modifying the belief or modifying the desire. Only *after* deciding which should be modified can the thinker apply methods of modification.

Such methods naturally include the "short-cuts" of wishful thinking and thinkful wishing. But I suggest that the method for deciding *which* to modify will contain within it suggestions for *how* to modify the chosen thought that do not rely on such shortcuts. When confronted with a tie between giving up the belief and giving up the desire, the thinker should look for tiebreakers. These tiebreakers could, and in practice *do*, vary case-by-case with the content of the thoughts in question. In the case of the pirate, content-wise nearby beliefs and desires include those about ships, robbery on the high seas, swords, parrots, and the like. On plausible views about mental content, my belief that I am not a pirate is in part *determined* by its place in the inferential network relative to these other beliefs; to believe I am not a pirate might just be something like the tendency to activate some combination of such other beliefs and images. Though nearby thoughts may not in this way come constitutively with the sparring propositional attitudes, at any rate the thinker should not have to look far for such tiebreakers. In some cases consultation of the nearby mental states will suggest that the desire should be modified or given up, and in some cases it will suggest that the belief should be modified or given up. In still other cases, I claim, examination of nearby states will result in the conclusion that *neither* should be

given up—the thinker should remain in apparent dissatisfaction.

For example: suppose I desire that I have some grapes, and I believe that I don't have some grapes. The content of these propositional attitudes is likely to trigger nearby thoughts about how grapes are found in stores or on vines; this in turn is likely to activate thoughts about how stores are places where you can obtain things using money, and about nearby stores in the area, the lack of vines in the area, and about what money I have. These nearby beliefs suggest my belief that I don't have grapes is easily modifiable; there is a way to modify this belief that I don't have grapes through intention formation and action, rather than direct manipulation. On the other hand where the mismatched p is "I'm a pirate", nearby beliefs include the one that pirates of the type I wish to be existed in the 17^{th} and 18^{th} century, which in turn is likely to trigger a belief about which century this is, a belief about the feasibility of going back in time, and so on. In this case it looks like the desire is what should be modified. And finally, suppose I am starving on a desert island (perhaps as a result of a failed pirate life), and there is in fact no food to be obtained anywhere. In this case it seems I am at an impasse; I should neither revise the desire to eat, nor the belief that there is nothing to eat. Instead, I should just remain in relative belief-desire incoherence.

Examining nearby content seems like a natural method for determining how to settle any given tug-of-war between belief and desire—it also happens to be roughly what we actually, typically, do. Section 2.2 and chapter III will give further details of this method, and show how to begin modeling it computationally. For now, we might still wonder: why not, once it's decided which of the belief and desire has to go, use one of the short-cut methods for changing it? And if it sometimes works, why not use it every time? For example, intention-formation and action may seem like a long way around for belief modification when compared to wishful thinking.

But it may turn out to be a preferable method given stubbornness in *spontaneous* beliefs about having grapes. This might also be a plausible explanation for why we *don't* typically do pure wishful thinking. Though one way beliefs arise is through inferential belief-forming mechanisms under something like conscious control, another way is spontaneous and not under our control; beliefs seem simply to filter up from our perceptions, from automatic (perhaps innate) inferential practices, and the like. These spontaneous beliefs—such as "I don't have grapes"—may persistently surface, causing lasting belief-desire tensions despite the wishfully believed "I have grapes." A cognitive mechanism capable of noticing this stubbornness in spontaneous beliefs may come quickly to prefer that in the case of something so easily attained through action, the intention-formation route is *easier* than the wishful thinking one at reducing long-term belief-desire tension. Thinkful wishing has similar problems. There are spontaneously-formed desires, too, and for humans a thinkfully wished desire not to eat will have a hard time in the face of spontaneous desires for food.

2.1.7 External constraints

But do we have reason to think that intelligent creatures will have, in general, such stubborn thoughts—ones that render wishful thinking and thinkful wishing ineffective? Yes. It is in virtue of being intelligent creatures at all that they will have such stubborn thoughts. Remember from chapter I that "creatures" are designed to perform functions of some kind. This design of the creature, it seems, will inevitably constrain their conations and cognitions; not every thought-forming mechanism is available to them.[22] Suppose I am right, for example, that feedback mechanisms for learning are responsible for cognitive-

[22]For an everyday example, our human design restricts us from forming cognitions about smells that a dog's design allows them to form. For an extreme but more conclusive example, a physically instantiated thinker could never possess a system capable of checking any reasonable number of atomic beliefs for truth-functional consistency, due to the enormous computational resources required. (For other examples see Stich (1990) pp151–154, which relies heavily on the oft-cited work of Cherniak (1986).)

conative pairs in the first place. These feedback mechanisms are then in the service of these creaturely functions. For a creature to develop a pure wishful thinking habit would be to short-circuit this feedback before it had a chance to influence cognitive dispositions toward its external aims, thereby ignoring the very reason the feedback mechanism was created in the first place. At least this part of the creature, essentially, would not be functioning well.

Wishful thinking and thinkful wishing look like objections to the internal pragmatic standard when we see apparent desire satisfaction as a creature's intrinsic cognitive aim, rather than merely instrumental in achieving external functions. Suppose, for comparison, I gave elaborate financial advice aimed toward the standard of "getting money". Someone might object: I could succeed at that standard just by making up my own currency and printing it in the basement. This *looks* like an objection to the proposed standard, intuitively, because printing my own currency would not do me the good that money is supposed to do me. But of course it is not really an objection, because the advice presumes in the first place a system where getting money is of further instrumental use.[23] Similarly, desires and beliefs are cognitive tools for the creature to perform its external function(s). The range of responses to mismatches in beliefs and desires is limited to what would serve those external functions—at least, in a creature that is functioning well.

This is not to say that wishful thinking is impossible for intelligent creatures. Of course we humans have amply demonstrated its possibility. But it is easy to forget that our own ability to do it is seriously constrained—we cannot simply believe that we're on a beach with a beer, no matter how much we might desire it. And there is reason to think that if we *do* manage to wishfully think pervasively, it is because we are not functioning well. Schizophrenics, for example, seem to be particularly adept wishful (or motivated) thinkers, to the point of hallucinations. Implicated in this mental malfunctioning is ex-

[23] You might object that it is not really "money" if I make it up and print it myself. But you could play the same game of conceptual legislation for "beliefs" formed directly from desires.

cessive dopamine, which seems to provide a kind of feedback to the brain for its own modulation.[24] Too much dopamine could, perhaps, "short-circuit" this feedback mechanism.

So the point is not that we couldn't possibly wishfully think; the point is that in virtue of being well-functioning intelligent creatures, our ability to wishfully think will be severely constrained, since it is practically guaranteed not to help fulfill the aims for which our thinking was designed. Make no mistake that the fulfillment of these aims is an *external* standard—it is an evaluation based on the performance of an external function, with which the thinker's own functioning can only contingently be correlated. I employ this external standard in defense of the internal standard of apparent desire satisfaction. Intuitions demand a reason to believe that wishful thinking (and thinkful wishing) will not in general be a rational (or "successful") way to think. But the brain-in-a-vat examples of chapter I make it plausible that wishful thinking is not always irrational. If the laws of nature were somehow such that a thinker's desires automatically brought about the desired state of affairs, wishful thinking *would* be a rational, successful way to think. The argument that wishful thinking is irrational must rely, therefore, at least in part on external factors like the way the world is, since it is only irrational in worlds like the one we figure to inhabit.

Meanwhile the fact that humans *do* engage in wishful thinking a great deal—and in thinkful wishing, for that matter—suggests that our brains do struggle to match cognitions to conations in whatever way possible, within the confines allowed them. A well-functioning brain in a "normal" environment will learn that wishful thinking does not work—which is why we think Bill from chapter I isn't so smart. (Indeed, Freudian theory has it that we start with a wishful thinking "pleasure principle" as infants, and only

[24]Dopamine seems to be crucially involved in reinforcement learning, for example; see *e.g.* O'Reilly and Munakata (2000) pp193–195.

eventually learn the "reality principle" through hard experience.[25])

So my response to the wishful thinking (and thinkful wishing) objection on behalf of the internal pragmatist, in the end, is partly that these mechanisms simply are not available to us, given reasonable assumptions about how the world is, and how we were designed by it. But that's not all there is to the story. Suppose these fallacious forms of reasoning *were* more generally available to us, say through some clever new form of neurosurgery that would make us ideal wishful thinkers (or thinkful wishers). Should we undergo the procedure, according to the internal pragmatist? No, of course not. But now the reason should be plainer. We shouldn't because we only aim to *believe* our desires are satisfied insofar as we aim to have our desires satisfied. The only way we have to judge whether they are or not is through our beliefs, of course. But our aim to have matches between our beliefs and desires only makes sense when we have reason to think that this aim would help bring about our external aims. I do not (in the first instance) want to believe that p—I just want that p. And as an intelligent creature I have internal mechanisms capable of learning in the service of this want, requiring (I have argued) internal representations of it and my relative success with respect to it. So I try in virtue of this want to bring my belief with regard to p in line with my desire that p. Our internal feedback functions, remember, were designed with the service of our external functions in mind. These functions, again, will provide external constraints on the kinds of thinking available to us. So the more complete version of the response, in summary, is that wishful thinking is not generally available to a creature that is functioning *well*, given facts about our environment and design.

[25]In many ways, Freud's theory of brain functioning is instructively comparable to my own—at least as I understand his theory, through articles like Hopkins (1998) (especially §7) and Glymour (1991). Whether this is a bad sign or a good sign I leave up to you.

2.2 Coherence

Wishful thinking and thinkful wishing looked like objections to internal pragmatic epistemology because they looked like ideal ways to achieve the proposed internal standard, while at the same time being intuitively repulsive methods for thinking—thereby reducing the standard to absurdity. But we have seen that they will not actually be good ways to achieve the proposed epistemic standard in an intelligent creature inhabiting an environment like ours. So, then, we should look for other cognitive mechanisms to reconcile mismatched beliefs and desires. According to the suggestion just hinted at so far, a creature that believes p and desires $\sim p$ can look to *other* "nearby" beliefs and desires in order to determine which should be modified, and how. Each of those belief-desire pairs, in turn, can be decided similarly. Such decisions must play off each other simultaneously. It looks like the system is doing better, then (according to the internal pragmatic standard) the better it balances all these belief-desire conflicts into a coherent state. And only some coherent states will be possible, given the external constraints on any intelligent creature. In this section I will provide further motivation for, and further elaboration of, this coherence approach.

2.2.1 Rational desires

The symmetry between wishful thinking and thinkful wishing reminds us that an intelligent creature, intuitively, should sometimes revise its goals as well as its beliefs. Typical epistemic pragmatists (like William Lycan, Nicholas Rescher, or Stephen Stich) seem to forget this intuition.[26] It's one thing for a pragmatist to say that given a thinker's goals, that thinker should be evaluated with respect to the achievement of those goals; it's quite another to say that the thinker should be evaluated on whether it comes up with good goals

[26] See *e.g.* Lycan (1988), Rescher (2001), and Stich (1990).

that then get achieved. Only the latter is consistent with the idea that part of good thinking involves choosing the right goals. It is no vice of a thinker's *belief*-forming mechanisms that it fails to achieve unattainable, crazy goals like becoming a pirate. And remember the stoic thinkful wisher, who desires things to be just as they are—even when, say, nuclear armageddon is moments away. This stoic shows it is also no virtue of belief-forming mechanisms to "achieve" easily attained goals. The core intuition is that desires can be rational or irrational, just as beliefs can. Intuitively a thinker is doing pragmatically well only if it comes up with rational goals, which it then (on the whole) achieves. So the evaluation of a creature's belief-forming mechanisms will depend in part on the success of the desire-forming mechanisms. But the evaluation of those will depend, in turn, on the belief-forming mechanisms, and so on.

Consider, for example, the formation of sub-goals. An intelligent creature must pick good goals in part by evaluating their feasibility. This will be determined in part by the estimated potential success of the sub-goals that it forms to achieve this goal. So goals will be picked in part by the strength of their sub-goals, but at the same time the sub-goals will be picked or discarded for their potential success at attaining the goal. When things are not working, it could be because the original goal is unrealistic, or because the sub-goal is ineffective.

These conflicts, in turn, will depend on conflicts between cognitions and conations. That, I have argued, is the basic impetus for change in cognitive disposition. The system would not need to worry about how to play off its conations against each other if all the correlating cognitions were matching. So take any cognition-conation pair resulting from a sub-function. For any mismatch between them, the system can continue to attempt improvement at the sub-aim, which would mean continuing to attempt to bring the internal representation of its success toward the goal state. Or, the system could decide that the

sub-aim is unfeasible, and demote it in favor of something else.

The resulting picture is a mess of small constraints, no one of which is decisive, and each of which the system struggles to satisfy in one way or another. This struggle to match cognitions and conations, again, is simply part being a sufficiently intelligent creature—and the more intelligence, it seems, the more complex such struggles will take place. Struggling to make one match may cause more mismatches elsewhere, or it may cause fewer. The challenge for the intelligent system is to fit all these together somehow, and come to an all-things-considered position both with respect to what it "believes" and with respect to what it "desires". The thinker generally cannot at any time satisfy *all* these mental constraints, so instead it is faced with a "multiple soft constraint" problem for playing them off each other, trying to violate as few as possible. Thagard has argued, in turn, that such a constraint satisfaction problem is just a more formal version of what we have always, in epistemology, called a problem of *coherence*.[27] The general notion should be familiar to epistemologists, but of course the proposed coherence is not typical *belief* coherence.[28] The proposal is that in virtue of being intelligent (in the way described) a creature will seek a *comprehensive* coherence, a coherence among *all* thoughts, cognitive and conative.

Now it may be that the best way to arrive at such comprehensive coherence is through separate attempts at purely cognitive coherence—and purely conative coherence, too. The creature is likely to attain more of its fixed aims given a coherent theory of the way the world is, rather than an inconsistent or disjointed theory. The more flexible and rich these aims, the more general and detailed the theory must be. And a coherent set of plans for altering the world to accord with the creature's conations also seems like a natural way to serve the comprehensive coherence at issue. (In chapter III we will see, in fact,

[27] See Thagard (2000) chapter 2.

[28] Of the type discussed by Laurence BonJour, say; BonJour (1985).

how specific proposals for calculating these local coherencies, like Thagard's ECHO and Thagard's and Elijah Millgram's DECO, might be integrated into such a comprehensive framework.)

The traditional alternative to justification from coherence, of course, is justification from *foundations*. In the case of internal pragmatic epistemology, it could be that there are foundational cognitions that support further cognitions inferred in the right way, and foundational *conations* that provide support for the conations derived from them. When it comes to specific conflicts between thoughts cognitive or conative, then whichever has the better foundational support will win (in a well-functioning thinker). And we have already seen, in discussing wishful thinking, that intelligent creatures will generally be severely constrained by their external functions in the mental mechanisms available to them. These external constraints may provide just the mental foundations required.

In fact, I think something like that foundational view is also right, though I will argue the externally imposed constraints provide *default* thoughts rather than *foundational* ones, and the ultimate calculation is still one of dynamic coherence rather than simple derivation. The view is thus actually a hybrid of coherentism and foundationalism—though closer to the coherentist side. The external constraints will make for only default (rather than incorrigible) thoughts, I will argue, due to the form these external constraints must take.

2.2.2 Intelligence and rich aims

Let's see how the externally imposed constraints are likely to work for an intelligent creature by imagining a detailed example. Suppose we want to build a creature that has the basic aim of keeping a public park clean. We outfit it with all sorts of perceptual tools that look, sniff, listen, feel, propriocept, and the like—and maybe a few tools capable of doing more direct chemical analyses of materials. We also outfit it with many kinds of grabbing,

moving, climbing, and digging tools, and whatever other potentially handy capacities we can think of. Once we managed to design those pretty well, the problem then becomes one of how to convert the inputs from the perceptual devices into appropriate outputs to the motor devices in a way that accomplishes its purposed aim. That is, we need to engineer its mental system.

As a first try, we could simply hardwire the motor outputs without any need to consult the perceptions. For example: "move five meters north by northeast, extend the arm this far, contract the pincers, move the pincers to the bag unit, release pincers, move twenty meters in new direction . . . " Naturally this will only work (the creature will only achieve its aim) insofar as the environment cooperates by having litter at the specified locations. Such a creature is unintelligent, for it is unable to adapt to an environment other than the one it is hardwired to expect. We can describe the hardwired aim not only as "keep the park clean", but also more simply as "move in these ways". This more specific way to characterize the aim tracks the lack of adaptability in trying to reach the more general aim.

So instead we take advantage of its perceptual capacities. Move around and look around, we tell it. And whenever you receive this exact combination of inputs, behave in this precise way. Naturally this would take a *lot* of programming on our part; we would have to teach it how any candy wrapper or aluminum can would look in any state of crumpledness or from any angle. Suppose we could manage, though. Suppose also we program this creature to keep a log of where and when it finds trash, and have it use this log to adjust its route accordingly. It now can *learn* to get more trash faster, keeping the park cleaner. Still, we could say its aim is "pick up stuff that's perceived in just this way" rather than "keep the park clean". For when new kinds of trash come along—a new candy bar, say—our creature will not pick it up. It cannot adapt to that kind of change in environment, and so is more likely to fail at the wider construal of its aim.

It would be still better if we could manage to provide for it a very broad understanding of litter. We would need to use abstract concepts like "artificial" and "left by a visitor". These concepts would have to be applied to determine what's litter according to some loose weighting, so that grass clippings left by a visitor would still count, and so would trash that had blown in on a wind. With concepts like "visitor" and "artificial", the creature could then even learn methods to prevent littering in the first place—it discovers, by accident, that wagging its claw at a visitor about to perform a certain type of action prevents the action, for example. With a rich enough conceptual library, the robot might learn to achieve its objective most effectively through initiating public service advertising campaigns, and so on.[29]

I have already noted that intelligence correlates well with adaptability, and that greater adaptability requires greater flexibility in learning. The point here is that such flexibility in learning requires the creature to have what we might call "rich" aims. Roughly speaking, rich aims will have a greater capacity for sub-aims according to various fine-grained environmental differences.

There is a similar theme running through work on naturalizing mental content. Think again of the simple thermostat from chapter I. As Dennett points out, its aim is better described as "keep the switch off only when the coil is so-long" than as "keep the room warm." His more complicated thermostat versions, however—the ones with multiple ways to detect room temperature, and multiple ways to regulate it—are more and more accurately describable as having the aim of keeping the room warm.

> ... as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If

[29] As will soon be obvious, this example was inspired by the complicated thermostat of Dennett (1981).

> you change its environment, it will *notice*, in effect, and make a change in its internal state in response.[30]

That sounds a lot like adaptability; that also, Dennett suggests, is what is required for a creature truly to represent its environment.

Fred Dretske has a similar hunch that there is an important tie between adaptability and rich mental content. He suggests that an approach like Dennett's will not work for naturalizing meaning, though, because it cannot handle the notorious disjunction problem for intuitive cases of misrepresentation:

> No matter how versatile a detection system we might design, no matter how many routes of informational access we might give an organism, the possibility will always exist of describing its function (and therefore the meaning$_f$ of its various states) as the detection of some highly disjunctive property of the proximal input. At least, this will always be possible *if* we have a determinate set of disjuncts to which we can retreat.[31]

Dretske follows this passage with a suggestion for fixing this problem, a suggestion especially interesting in the context of my own view:

> Suppose, however, that we have a system capable of some form of associative learning. ... We now have a cognitive mechanism that not only transforms a variety of different sensory inputs ... into *one* output-determining state ..., but is capable of modifying the character of this many-one mapping over time. ... A system at this level of complexity, having not only multiple channels of access to what it needs to know about, but the resources for expanding

[30] Dennett (1981) p235.

[31] Dretske (1986) p338; 'meaning$_f$' is roughly *functional* meaning along teleosemantic lines. Dretske is not explicitly responding to Dennett here.

> its information-gathering resources, possesses, I submit, a genuine power of misrepresentation.[32]

If this is right, then to have truly rich content—content capable of genuine misrepresentation, for example—requires at least primitive learning mechanisms. The rich content of the aims of an intelligent system consists, in part, of the wide variety of sub-aims the system can pursue in achieving the aim.

2.2.3 Defaults and foundherence

Imagine, then, our park-cleaner has *clean park* as its *basic aim*—an aim ultimately responsible, in some way, for its design (in this case, through our own intentions to have a robot to clean the park). This conceptually rich goal is constituted, we have seen, by many smaller sub-aims it is hardwired to seek. "Scour the park, look over the whole thing carefully! Pick up stuff that meets these *trash* criteria! Keep yourself in working order! Prevent stuff that meets these *littering* criteria!" Let's call the aims that came hardwired for attaining the basic aim *fixed aims*.[33] These then each have sub-aims of their own, internal feedback mechanisms redesigning the creature's own cognitive system toward the matching of its cognitions to its higher conations. "Go south! Go north! Practice grabbing! Test the hypothesis that sample X is trash! Plan a new route! Scold those litterbugs! Fix your left rotor!" A genuinely intelligent creature will teem with such potential sub-aims. Each of these either constitutes a little feedback mechanism in turn, with its own implicit conation-cognition pair, or else bottoms out in a basic function (such as a motor impulse).

Unlike the fixed aims, though, the sub-aim conations are not immutable. We were clever enough to design the robot so that they can be learned and adjusted, under the

[32] Dretske (1986) p338.

[33] Though these are only default aims, as we will see, they are "fixed" insofar as they cannot be undefaulted, so to speak. In terms of the computational model to come later, at least some of their activating constraints cannot be adjusted by the system (namely, the constraints with special element @).

guidance of the basic aim, as constituted by the fixed aims. As suggested before, there are two ways these conations might change: first, the creature may learn that even when it succeeds (subjectively) in matching the cognition to the conation, the higher function in whose service it was formed fares no better. On the other hand, the creature may learn that no amount of learning, and no sub-sub-aims, will bring about that sub-aim's success. The cognition and the conation simply refuse to match. In both cases the sub-aim does not ultimately contribute to the higher aim, and in both cases a smart system will as a result inhibit the functioning of this sub-aim.

If such conflicts always had some easy resolution according to feedback from the aims above, there would be no difficulty in calculating what to do. Similarly, an instrumentalist about practical reason has little difficulty explaining which goals are good or bad. Irrational desires are those deriving from irrational *beliefs* about what means would facilitate the higher ends. The highest ends, then, are taken as intrinsic, immutable, and rational—as foundational, in other words. In our example, the creature need merely determine through feedback which of several possibilities best achieves the aim of *clean park*.

But the matter is not so simple, because it seems any creature with rich aims will have a variety of fixed aims that may themselves lead to conflicts. For example: constituting its basic aim of a clean park, our park cleaner has fixed aims both to take care of itself, and to pick up trash that it sees. Reaching way over a cliff to retrieve some garbage will be a good means to one, but not the other. The creature must somehow balance these aims against each other, and since they are fixed, there is no higher internal aim to which it can appeal in deciding. With several such potentially conflicting fixed aims for any potential decision, the situation becomes that much more complicated. Not all of even the fixed aims can be served. Some will have to be violated for the sake of the others. In other words, the problem is one of maximizing soft constraint satisfaction—one of coherence.

Again this coherence is not free to settle into any old configuration, as we have seen; the designed aim of the creature imposes external constraints in the form of these fixed functions. In effect it will have *default* conations hardwired, constituting its rich aim as an intelligent creature. And it will be hardwired to represent its success at these conations in default ways—the creature will simply receive basic signals from its sensory periphery, for example, and those will be thrown in the coherence hopper as default cognitions. Though any one default conation or cognition might be overridden (as our robot's overriding the default "get all trash" directive in favor of the "preserve self" directive), it would only be in order to satisfy enough of the *other* default cognitions and conations. Not *all* of them could be overridden, because (I claim) it is in the nature of the intelligent system to satisfy as many as possible in pursuit of its external function(s).[34]

I should emphasize that the result is not a pure coherentist position. Some thoughts—the default cognitions and conations—are more important to satisfy than others. Put in terms of traditional epistemology, these thoughts contribute more toward overall justification when accepted by the system. The external constraints on the coherence thus provide some justificational *foundation*, if you like, for the rest of the thoughts. But of course the position is not a pure foundationalist one, either. The derivative thoughts have no direct inferential foundation, for one thing, but rather result from a balance of the others. And any of these default thoughts is in effect defeasible, rather than truly foundational. The proposal is therefore a hybrid position of the type Susan Haack calls "foundherentist".[35] It may seem to be a hopelessly wishy-washy position, to have "default" thoughts that are *sorta* foundational. But as we will see, Thagard has shown how there is a natural, perfectly respectable algorithm for computing such foundherentist problems—an algorithm that for

[34]Here and elsewhere to "satisfy" a conation should be read in the subjective sense, as getting a matching cognition.

[35]Haack (1993).

example our own brains seem eminently capable of computing.[36] My proposal builds on this computational construal of foundherence.

2.2.4 Coherence and humans

This is all a fine story, perhaps, about abstract intelligent creatures. But is it at all plausible for, say, the intelligent creatures most familiar to us—namely, *us*? In the case of humans, we were designed by natural selection to pass on our genetic material. In the service of this aim, it seems we were given a mass of fixed aims—probably ones like avoid pain, get food, have sex, gain affection, give affection, and so on. We are so remarkably flexible at attaining these fixed aims that it is often at the expense of the evolutionary aim for which they were initially designed. We can keep seeking food beyond what our health (and thus genetic fitness) can sustain. We can have sex without having kids. And so on. Similarly, the balance of our park-cleaning robot's fixed aims could end up violating the main aim for which it was designed. It may find, for example, that the best way to keep the park clean is to keep all visitors out of the park—by any horrific, *Terminator*-like means necessary. Oops, we forgot to give the robot a rich enough notion of "park" to include allowing visitors. Well, no designer is perfect, after all—not even Mother Nature.

And, again, humanly fixed aims like getting food can lose out to a complex of other fixed aims, as for example Mahatma Gandhi demonstrated. Of course I do not claim that "freeing India" was a competing *fixed* aim for Gandhi. But it did arise from a complex of other such fixed aims in Gandhi, such as perhaps caring for other people. It must have—from where else could it have come? (Incidentally I don't suggest that Gandhi was being *selfish* in starving. Probably some of our fixed aims are altruistic; or if not, it is probably rational to learn altruistic aims in their service.) The point is: we are not slaves to each of our fixed aims, but we are not completely free of them either. Our humanly desires are not

[36] See especially Thagard (2000).

sui generis—they are either fixed aims or derived somehow from a balance of them. Of course a human may come to have irrational sub-aims, such as serial killing. Intuitively, those humans are sick or malfunctioning in some way; they have failed to reach the right conclusions about what sub-aims to form. They are in bad belief-desire incoherence; they are also, we think, desperately unhappy.

Naturally we humans also have fixed cognitions, and they too are *default* constraints that may not be met. We just receive a lot of sense-data. Any packet of sense-data can be overridden; we can think ourselves out of optical illusions, for example. A pilot banking in the clouds will eventually become convinced that she is flying level—at least until she learns to trust the plane's artificial horizon monitor, telling her otherwise. But we are not free to ignore all these cognitions, either. We have to take most of them at face value. A common starting point of internal epistemology is "trust, *ceteris paribus*, your senses."[37] Fine: but *why*? I would say we cannot *help* but do this. It is part of being an intelligent creature that we will form cognitive representations of the world through mechanisms that are simply given to us. They and our basic conations form the original substance of our thoughts. The more interesting question is when and why we shouldn't, and how and on what grounds we make such decisions. I have argued that the ultimate standard for such decisions is a general coherence between conations and cognitions, within the bounds of externally imposed constraints—making for a more specific, and more easily naturalized, version of internal pragmatic epistemology.

[37] See *e.g.* Pollock and Cruz (1999), or Wedgwood (2002).

CHAPTER III

COMPUTATIONAL EPISTEMOLOGY

As I have emphasized throughout, one of the best features of this account is that it has real potential for a computational model. In fact the epistemic considerations thus far amount to a set of specifications from which such a computational model of intelligence might arise. In this chapter I will first detail these specifications. Then I will provide an algorithm schema for meeting them, and a realistic example of an architecture for computing that type of algorithm. This is the last piece of the "computational internal pragmatic coherence epistemology" that I promised.

With the whole view finally in place, I will then take some time to examine its advantages, its implications, and its suggestions for further research.

3.1 Belief-desire coherence (BDC)

This section will finally attempt to make the promised link from philosophy to cognitive science, by providing an algorithm and architecture to help realize the epistemology of the previous sections.

Really, though, I should say that the order of explanation here is a bit artificial. In some ways the existence of the algorithm argues for the philosophy rather than vice-versa. And in some ways the existence of my chosen architecture argues for the specifications for the

algorithm, rather than vice-versa. And in general, I see the various aspects of the view as supporting each other mutually, in a neat *coherence*, rather than following one upon the other with only unidirectional support.

3.1.1 Specifications

At any rate, to summarize the foregoing philosophy, we now have reason to think that the mental system of a sufficiently intelligent creature must have the following kind of computational capacity:

- It must be able to *learn* better performance at a task based on an internally available measure (thereby possessing conations and cognitions);
- That measure should be one of a certain kind of *coherence* among all the conations and cognitions the creature thus formed;
- That coherence should allow for *default* thoughts (both conative and cognitive) that have a kind of defeasible precedence.

To this list I will add two more desiderata that we have not yet explicitly discussed.

- The system should allow for *degrees* of acceptance (and rejection) of the conations and cognitions;
- The system should be *physically realistic*, able to compute the coherence problem in a reasonable amount of space and time.

The latter should be uncontroversial for anyone looking to naturalize intelligence by showing how it could arise in an ordinary physical creature. The former has two, independent motivations: first, suspicion on philosophical grounds that beliefs, desires and such are

not, in the final analysis, all-or-nothing affairs. I will not defend or elaborate this suspicion, but trust that enough others share it, or anyway are willing to tolerate it. Second, degrees of acceptance are important for systems that learn through feedback mechanisms. As Randall O'Reilly and Yuko Munakata put it:

> . . . graded changes allow the system to try out various new ideas (ways of processing things), and get some kind of graded, proportional indication of how these changes affect processing. By exploring lots of little changes, the system can evaluate and strengthen those that improve performance, while abandoning those that do not.[1]

In other words, it is important when trying different approaches to improving at a task to be able to make fine-grained adjustments in behavior, both cognitive and external. (The flip side of this coin is *graceful degradation*: a system based on degrees of acceptance that fails or malfunctions in some way can do so partially, without disastrous consequences cascading throughout the system.)

Perhaps talk of the system "trying" new ideas and the like, though, sounds too strange at this point. How does a system of wires or neurons "try" different methods, and "decide" to prefer these mechanisms over those? It may seem to require a cranial homunculus, testing various techniques and picking its favorite. If so, it is unclear what work the standard is doing for either artificial intelligence or normative epistemology—since designing such homunculi for machines only pushes the problem back a step, and since we presumably have no such homunculi ourselves. How can the brain of a park-cleaning robot or a human being make its own functioning better (according to the standard of comprehensive coherence) without a homunculus? We turn now to cognitive science for the beginning of an answer.

[1] O'Reilly and Munakata (2000) p18.

3.1.2 Machine learning

Modeling learning in a machine is certainly a tricky problem for any cognitive theorist, but there has been progress along several different tracks: for example, the number-theoretic track based on seminal work by Hilary Putnam and others. Closely related to the number-theoretic track is the Bayesian track, perhaps most actively pursued now by Clark Glymour. Or there is the less unified track examining prospects for goal-driven learning.[2] For reasons that will soon become clear, I will look briefly at learning from a *connectionist*, or artificial neural network, standpoint. The point for now is merely to provide an example of how a machine might learn.

Patricia Churchland and Terrence Sejnowski provide a taxonomy of learning algorithms for neural networks in Churchland and Sejnowski (1992). Networks, they say, can learn through some combination of external and internal feedback; in the former case something outside the system provides information about the quality of the answer, while in the latter case the system can test itself. For clarity Churchland and Sejnowski call nets with external feedback mechanisms "supervised", and nets with internal feedback systems "monitored." (The former term has caught on, and usually systems with only internal feedback are called "unsupervised".) The notion of internal feedback for a network may seem mysterious to some—and since it is crucial to the project, it is worth some explanation.

> Consider, for example, a net required to learn to predict the next input. Assume it gets no external feedback, but it does use its previous inputs to make its predictions. When the next input enters, the net may be able to use the discrepancy between the predicted input and the actual input to get a measure of error, which it can then use to improve its next prediction.

[2]For the number-theoretic version see Jain et al. (1999) and Kelly (1996); for a Bayesian version see Spirtes et al. (2001); for goal-driven versions see Ram and Leake (1995).

This is an example of a system with internal but no external feedback; the learning is "unsupervised" but "monitored". "More generally," Churchland and Sejnowski say, "there may be internal measures of consistency *or coherence* that can also be internally monitored and used in improving the internal representation."[3]

Networks other than the most simple ones are good at extracting what Churchland and Sejnowski call "higher-order information" from large inputs. (For example, the visual cortex extracts information like surface boundaries from retinal cell impulses.) There are generally two tasks in extracting such "higher-order information": first, to see what kinds of patterns and correlations are in the information; second, to sort through and represent only the patterns of *interest* out of all the possible ones. "The information for this last task," they say,

> cannot be garnered from inside the net itself, but must be provided from the outside. The division of labor in a net with hidden units looks like this: unsupervised learning is very good at finding combinations but cannot know which subset to "care" about; supervised learning can be given criteria to segregate a "useful" subset of patterns, but it is less efficient in searching out the basic combinations.[4]

For many systems, then, a combination of internal and external feedback is ideal.

Plausibly the human brain is both a supervised and monitored non-artificial neural network. Evolution provided the *design* of the human brain with sloppy but (over vast time) effective external feedback. It is thanks to evolution that we care about some higher-order information—like surface boundaries, or who loves us—and not about other information, such as infrared spectra or who has detached earlobes. In this sense evolution provided us

[3]Churchland and Sejnowski (1992) p97, emphasis in the last quotation is of course mine.
[4]Churchland and Sejnowski (1992) p99.

with basic things to care about, understood as fixed conations.

The problem for a net with an internal feedback mechanism is to adjust the weights between the nodes of the network in a way that best facilitates minimizing error. This is a difficult problem.

> Finding a suitable weight-change rule looks really tough, because not only are the units *hidden*, but they may be *nonlinear*, so trial and error is hopeless, and no decision procedure, apart from exhaustive search, exists for solving this problem in general.[5]

Still, for particular kinds of nets, even quite complicated ones, weight-adjusting rules exist:

> It is now clear that there are many possible solutions to the weight-adjusting problem in a net with hidden (and possibly nonlinear) units, and other solutions may draw on nets with a different architecture and with different dynamics. Thus nets may have continuous valued units, the output function for a unit may have complex nonlinearities, connections between units need not be symmetric, and the network may have more interesting dynamics, such as limit cycles and constrained trajectories.[6]

Of course there has been much development in this direction since this quotation was first published, some ten years ago. Neural networks are steadily capable of solving more and more complex weight-adjustment problems, which for our purposes are simply learning problems (as we will see).

3.1.3 Formal coherence

Now to meet the specifications from section 3.1.1. Let's start with a formal construal of coherence problems generally. I want to start here partly because it is easy: Paul Thagard

[5] Churchland and Sejnowski (1992) p100.
[6] Churchland and Sejnowski (1992) p102.

has already done this work for us. In Thagard (2000) he argues persuasively that philosophical coherence is best understood as a constraint satisfaction problem (what I have also heard called a "multiple soft constraint" problem). Here is a formal characterization based on Thagard's.[7]

- Let E be the set of elements e_i that must be brought into coherence.
- Let $a : E \rightarrow [-1, 1]$. This represents the degree of *acceptance* (for positive numbers) or *rejection* (for negative numbers) of any particular element.
- Let $c : E \times E \rightarrow [-1, 1]$. These are the *constraints* that must be satisfied among the elements. Typically c is such that $c(x, y) = c(y, x)$.
- Let $C = \sum_j \sum_i c(e_i, e_j)a(e_i)a(e_j)$, representing the *coherence* of the system. A positive constraint between elements will contribute to coherence when their acceptance values are near; a negative constraint between two elements will contribute to coherence when their acceptance values are far.[8]

Then the *coherence problem* is to find the acceptance function(s) a that maximize the coherence C given the constraint function c. Furthermore, I propose that there is an associated *learning problem* for systems that can learn: namely how to gain better *potential* coherence through adjustments to the constraint function c. The system's learning consists,

[7]Thagard's own formal characterization does not allow for *degrees* of acceptance and rejection until he puts it in a connectionist framework. But as the specifications make clear, I think the matter of degree should be understood as part of the abstract problem type. It is surely a desideratum of the general algorithm, independent of architectural implementation. (But listing this as a desired feature ahead of time does make a connectionist architecture all the more appealing.) For Thagard's own all-or-nothing characterization of coherence problems see Thagard (2000) p18.

[8]The standard measure of what is called "harmony" in a constraint satisfaction problem would in this case be $H = 1/2 \sum_j \sum_i c(e_i, e_j)a(e_i)a(e_j)$ (see O'Reilly and Munakata (2000) p107); I ignore the one-half term for the double-summing. Incidentally, contrast C with a measure like $C' = -\sum_j \sum_i c(e_i, e_j)|a(e_i) - a(e_j)|$. C' prefers acceptance functions the better they are able to bring together the values of positively-constrained elements, and drive apart the values of negatively-constrained elements. But C will give additional preference to systems that give positively-constrained elements the same sign, and negatively-constrained elements opposite sign. It also prefers systems with low entropy—systems with fewer acceptance values near zero.

then, in its ability to change the coherence problem it computes. An intelligent creature is functioning well to the extent it has reached coherence between its cognitions and its conations, and is learning well to the extent that it has changed its functional nature in order to improve its potential for such overall coherence. (Learning well, then, is *part* of functioning well.)

Now that we have a computationally respectable understanding of coherence, let's specify the specific coherence problem involved in learning for intelligent creatures.[9]

3.1.4 The belief-desire coherence algorithm

We will start with an algorithm specified at the level of propositional attitudes. (Later we will see how it might be extended to a sub-propositional level.) Call this the *belief-desire coherence algorithm*, or BDCA for short. I should emphasize this is not strictly an algorithm so much as an algorithm schema; the details of the algorithm would depend on the creature.

First, consider all the propositionally expressed internal aims of the creature—that is, aims that arise from a conative-cognitive representation pair about some proposition p. There will be a desire with respect to p, or d_p, and a belief with respect to p, or b_p.[10] Let all these *thoughts*, conative and cognitive, be the set of elements T to be brought into coherence. Let c be a commutative function where $c(d_p, b_p) = 1$ for each p. With such constraints, the system will be in greater coherence when more aims are matched. That is, C will be greater the more the acceptance function assigns both parts of each belief-desire

[9]Computationally "respectable" is not to be read as synonymous with "tractable"; actually, Karsten Verbeurgt has shown in Thagard and Verbeurgt (1998) that coherence problems are NP-hard and therefore overwhelmingly likely to be *in*tractable, even before countenancing *degrees* of acceptance as in my version. Still, as Thagard puts it, there are "several effective approximation algorithms" (Thagard (2000) p15), most notably a connectionist one.

[10]Strictly speaking, d_p is not necessarily a desire, but only some conation; similarly b_p is really some cognition that need not be a belief. (See section 3.2.3 for a discussion of non-belief cognitive propositional attitudes and the like.) To call them all "beliefs and desires" here is largely for convenience.

pair either strong acceptance or strong rejection. These constraints are fixed.

Add to the set T a special element, @. This element represents the external constraints imposed on the system by the actual world. Let the coherence problem range only over acceptance functions a that map this special element to 1. This special element determines the *default* thoughts. It is always highly accepted, and positive constraints with some chosen beliefs and desires make it likely, but not guaranteed, that those will be accepted, too. (Similarly a negative constraint with @ makes a thought default rejected.) If there is enough conflict in the rest of the system, though, any default thought can be overruled—as per our specifications. For each default accepted thought t let $c(@, t) = 1$, and for each default rejected thought t let $c(@, t) = -1$. These constraints, too, are fixed.

Finally, specify constraints between the thoughts according to propositional content. For example, contradictory beliefs should have a strong negative constraint between them. A belief with an explanatory or evidential relation to another belief should have a positive constraint between them. A desire that if satisfied would be a means to a further desire should have a strong positive constraint between them. A belief that if true would make some desire impossible should have a negative constraint between them. And so on.[11] These non-default constraints can vary to some degree as the system learns (how freely depends on the basic design of the system). The default cognitions should be propositionally "basic" ones—something like early Wittgensteinian atoms, or Russellian sense-data. The default conations for intelligent creatures, though, should be the opposite—highly abstract and complex, or what section 2.2.2 called "rich".

Admittedly, that last step is very hand-wavy. It is an unfortunate byproduct of leaving the specification at the level of folk psychology and propositions, neither of which is very

[11] As we will see in section 3.2.1, I have in mind connections of the type made in Thagard's DECO and ECHO models; again see Thagard (2000) for an overview. DECO is Thagard's model of *deliberative* coherence, developed jointly with Elijah Millgram; see *e.g.* Millgram and Thagard (1996).

well understood when it comes to physical implementation. I will suggest a step toward remedying this in a moment; for now let's review the intuition behind the BDCA. It is first a subjective *measure* of the system's success at balancing its default desires according to what is reported via its default beliefs. But second, it is also a force toward which the rest of the system can learn. Fixing the constraints between belief-desire pairs at 1 demands that when the system comes to reassign the other constraints, it will be towards greater belief-desire coherence. Some of these intermediate constraints the system may not be able to adjust, but others it can. To the extent it is able to adjust these other constraints, I claim, the system is capable of learning.

3.1.5 The architecture, and a toy example

Now let's look at the "physically realistic" specification for our algorithm. The question here is whether BDCA can be computed by a physical system using reasonable resources. Well, constraint satisfaction problems are the bread-and-butter of artificial neural networks, which are efficient estimators of solutions to such problems.[12] There is also, as we have seen, a great deal of work on how such connectionist networks can learn to adjust their weights according to internally-detectable error functions. And mapping a coherence problem onto a neural network is easy: let the nodes of the network represent the elements, let the constraints of the problem determine the weights between the nodes, and let the network calculate the activation function a by iterating the network through different activation states according to a standard updating rule.[13] We also know that some kind of parallel distributed processing roughly along connectionist lines is just the kind of thing that, for example, the *real* neural networks of the human brain manage to do all the time.

[12] Especially in the "degree" formulation of coherence problems, as noted; but see chapter 2 of Thagard (2000) for more of an argument to this effect.

[13] The simplest of which is $e_j = \sum_i c(i,j)e_i$; often this sum is fed into a sigmoidal function that limits the values asymptotically to an interval like our $[-1, 1]$.

Thus our algorithm seems like a cognitively realistic computation for a physical system to approximate, and within reasonable time constraints.[14]

With a connectionist picture in the background, we can also begin to see how this feedback mechanism might integrate into the rest of a mental system, without needing to rely on the mysterious primitive connections forged via propositional content. In a very simple, unintelligent creature, inputs (in the form of sensory stimulations) get mapped directly to outputs (in the form of motor stimulations). In the case of more complex but still unintelligent creatures, the inputs are sorted in a pre-wired way for patterns, or higher-order information—such as sorting visual stimuli for surface boundaries, and then sorting those for boundaries typical of predators. (Again, artificial neural networks model such extraction of higher-order information very well.) These patterns of stimuli will directly cause patterns of output, such as the pattern of motor stimulation required for moving away from the predator (given the position of the predator and such). In a complex creature capable of learning, coaching from internal feedback can shift the way the higher-level information is extracted from the senses, and shift the way low-level motor coordination results from high-level desires. In a network model, "credit" and "blame" for internally-detected successes and failures can filter down recursively through the layers of representational patterns. In this way the internal standard of aim-matching can provide feedback throughout the mental system.

To get a sense of how this works, consider a toy example. Imagine a creature with a mental system like this: first, it has a layer of sensory input nodes. Let these feed a second layer to extract patterns from these inputs, and let a third layer extract yet higher-

[14]There is also active, empirically-backed speculation about how the constraint satisfaction form of learning takes place in humans. For example, O'Reilly and Munakata (2000) p420 mentions that the anterior cingulate cortex seems to have much to do with learning, acting somehow like the "adaptive critic" feedback mechanism for reinforcement modeling in networks. Significantly, the anterior cingulate cortex has also been implicated in processing emotions, which I am inclined to read as various types of belief-desire incoherence; see section 3.2.4.

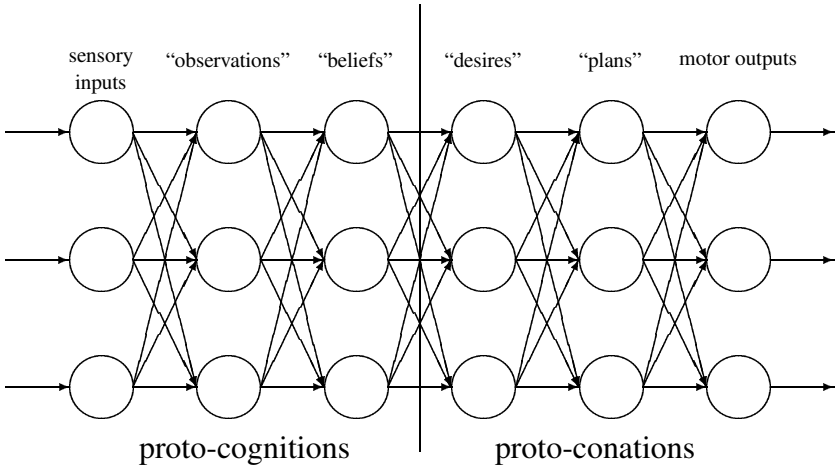

Figure 3.1: A somewhat sophisticated, but unintelligent, creature.

order patterns from those patterns. Think of the periphery layer as raw data, the second as somewhat theory-tainted observations, and the third as abstract beliefs and theories about how things are. These are the proto-cognitive layers.[15] Similarly let the creature have three proto-conative layers—one to represent, roughly, the abstract desires resulting from fixed aims, the layer below it for plans to bring them about, and the bottom layer for motor output. (Of course each of these layers is likely to represent many many levels of computation in any actual creature.)

Now imagine rigging these two systems together as in figure 3.1, hooking up the high-level proto-cognitions to the high-level proto-conations, so that the latter can take advantage of the former in fixed ways. (We assume that they will mesh well; that is, the kinds of patterns that the high last cognitive layer sorts out are the kind that the high conative layer is interested in.) The result is a somewhat sophisticated creature, able to react to various patterns in its environment in fairly complex ways—but still rather unintelligent by our standard, since its computations are hardwired.

[15] I call them "*proto*-cognitive" because in a creature that cannot learn I suspect there is no principled distinction between cognitions and conations (as the figure illustrates); see section 1.3.

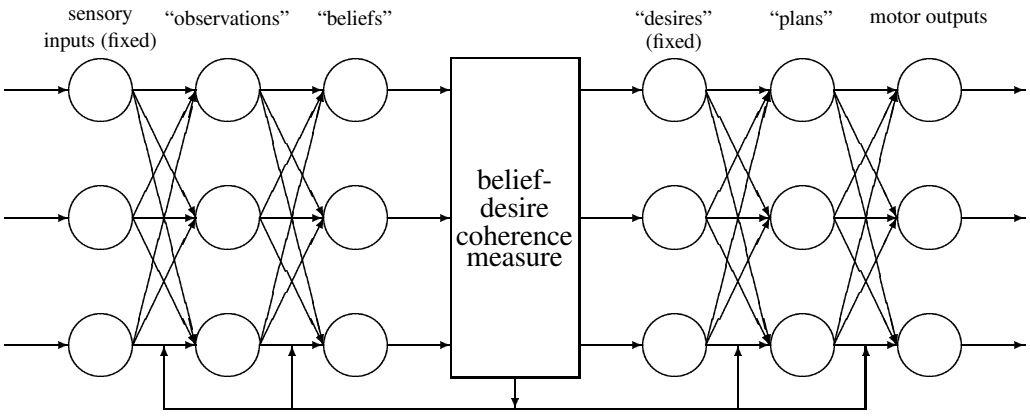

Figure 3.2: A toy example of the BDCA.

So finally, to give it the ability to *learn*, there must be ways to change the weights between nodes in the network. That allows the creature to compute different functions from input to output; different patterns can emerge from the lower-level cognitions, and different kinds of actions can result from the higher-level conations. To change its computational dispositions for the better requires an internal measure of how the creature is doing at matching its aims. So we place fixed weights at the juncture between its cognitions and conations, demanding matches. Error from mismatches can propagate back recursively down through the layers, and the weights of the network that are flexible will shift in its favor. The toy result would look something like figure 3.2.

In such a coherence and learning problem, the low-level, sensory input will receive the default activation on the cognitive side. On the conative side, it is the high-level conations that would receive the stubborn, default activations. These conative defaults artificially represent the wiring bias the system would have, in virtue of being a creature at all, toward reacting with certain types of actions given certain cognitive circumstances (like running in the presence of a predator). Together the cognitive and conative defaults represent the

role the external world has to play in its thinking—in both the inputs the world gives the creature, and the fixed aims the world has designed the creature to have.

3.1.6 Next steps

Naturally this characterization, too, is approximate and artificial. For one thing, there may not be any sharp boundary between cognitions and conations even in creatures that can learn. It may be more like a matter of degree. When surface boundaries are extracted from visual input, that already represents some mix of the creature's interests with its cognitive representations, since the creature has reason from higher up to pick out these boundaries. And when a conation involving patterns of movement (like "flee") translates to motor impulses, that represents in part a means-end cognition about how to flee. To reflect this, the model of the aim-matching feedback mechanism would have to be adjusted accordingly. On this conception of the learning process, there is no one black box monitoring the creature's highest-level thoughts. Instead we would have a network of many layers between input and output, with no sharp boundary between cognitive and conative layers, and with lots of local and global monitoring doing the work of the one big black box. In this model conations of the creature are perhaps best represented simply by fixed bounds within which the constraint function c can vary. And wherever there is a monitor for computational improvement (weight adjustment), there we can say exists an implicit cognition-conation pair—the monitoring mechanisms adjust the creature's computation toward achieving activation states that the creature *wants*. Or I should rather say: the creature wants those activation states in virtue of its function to adjust the computation toward their achievement.

For another thing, when it comes to modeling intelligence on the level of human brains, there are many factors to consider that are not easily reflected in an artificial neural net-

work. Perhaps most obviously, the sheer size of the required network is intimidating, if the human brain is any indication. The 18 nodes in our toy creature (plus however many in the feedback mechanism) are pretty paltry compared to our own 100 billion neurons. Also the network of an intelligent creature will require a great deal of recurrent (backwards-reaching) paths in order to allow the creature to model sensory and motor processes through time, and to disambiguate degraded input.[16] The effect is another kind of feedback to the system, and its interaction with the proposed BDC feedback is unclear. Finally, this model neglects important computational aspects of biological creatures. For example, it replaces the discrete spiking function of neurons with a continuous-valued output. And perhaps most importantly, it ignores (or assumes) the effects of neurotransmitters and other chemicals. Dopamine, for example, appears to be instrumental in adjusting synaptic weighting between neurons, and thus (many suspect) may be crucial to the learning process.[17]

3.2 Advantages and further issues

Still, as a step toward integrating abstract epistemology and practical cognitive science, BDC shows promise. Already we have seen the theory integrate crucial intuitions of internal epistemology, intelligence, learning, coherence, and the computational theory of mind. But, as I hope to show in this section, its potential does not stop there.

3.2.1 DECO, ECHO, and BDC

For one thing, BDC can integrate smoothly with some of the work Thagard and others have started in computational epistemology. For example, he and Elijah Millgram have come up with a computational model called "DECO"—based on Thagard's earlier work

[16] See Churchland (1995) chapter 5 for a fun, accessible discussion.

[17] See p50 and the reference in footnote 24 on the role of dopamine; for more on the general topic of chemicals and mental computation see *e.g.* Thagard (2002).

modeling explanatory coherence—to show how decisions might be made through a kind of *deliberative coherence*. They propose deliberative coherence as a constraint satisfaction problem where:

- the *elements* are deliberative "factors" such as goals, subgoals, and actions (among which DECO does not distinguish);
- the *constraints* consist of positive, symmetric constraints between factors that facilitate the accomplishment of other factors, and negative constraints between factors that are difficult to realize together;
- the *interpretation* of acceptance in the model is a decision on the part of the agent to adopt a coherent set of factors—that is, to form a holistic plan of action.[18]

DECO has several advantages. First, it captures an approach to instrumental reasoning. Second, it allows for reconsideration of ends; they can be outweighed if there are enough negative constraints with other deliberative factors that are highly activated. Thus while incorporating instrumental reasoning, it allows that sometimes agents deliberate about ends, too. Perhaps most obviously, it provides the beginning of a computational model for problems in practical reasoning.

But DECO takes its constraints, such as the facilitation relations, as primitive. As a result, it is incapable of learning on its own. Plugging DECO as a starting set of constraints into a BDC system would allow the system to learn better plan-formation over time. Also, DECO's authors recognize that *beliefs* have a heavy part to play in deliberative coherence. At the very least, a thinker's beliefs should affect what kind of facilitation relations hold between conative states. Therefore the authors provide a "principle of judgment", stating that "facilitation and competition relations can depend on coherence with judgments about

[18] See Thagard and Millgram (1995); Millgram and Thagard (1996); Thagard (2000, 2001).

the acceptability of factual beliefs."[19] Just how this dependence works is left out of DECO, but it seems a crucial issue. BDC could potentially explain the formation of these facilitation constraints as indirectly conducive to satisfying its own constraints, and could also explain why facilitation relations are formed based upon antecedent cognitions. Basically, it *works* to do so—where belief-desire coherence provides the final internal standard for what "works".

Here is another problem DECO has on its own, but that might be solved by integrating it into a BDC system. Thagard and Millgram see their proposal as an alternative to classical decision theory. Among the advantages touted, DECO explains the preferences that decision theory leaves primitive, allows for the backwards formation of goals from subgoals (such as deciding to adopt the goal of running a marathon from a desire to run every day), may model and explain irrationality with respect to the axioms of decision theory, and against a naive instrumentalist it allows for deliberation about the preferences involved.[20] Still, DECO cannot diverge too far from utility theory and remain plausible. For suppose DECO endorses plans that are highly "coherent" in terms of a thick web of reinforced facilitation constraints, but that bring about less of what the agents want. Then DECO would presumably lose much of its normative force, for why should an agent prefer the *coherent* plan to the fruitful one? In some sense the more coherent plan—the plan that best balances deep spontaneous desires against possibilities for achieving them—must *be* the one with more utility to it. Such an outcome would be much better assured if the constraints were adjusted toward exactly such a purpose. When subjective utility is construed as belief-desire coherence, and when deliberative coherence constraints are formed in the service of this belief-desire coherence, the result is one neat package.

[19] See any of Thagard and Millgram (1995) p442, Millgram and Thagard (1996) p73, Thagard (2000) p128 for this quotation.

[20] See Thagard and Millgram (1995) pp449–452 for more details.

The proper functioning of DECO under a BDC learning mechanism will depend in part, of course, on how well the system as a whole forms beliefs, so we should take a look at how to form those. Again, Thagard's work is helpful; we have already seen in section 1.2.5 how his ECHO system models *explanatory* coherence. With the computational take on coherence under our belt, we can look at ECHO more detail.

- the *elements* are cognitions;
- the *constraints* consist of positive, symmetric constraints between propositions where explanatory relations hold or analogical mappings exist, and negative constraints between contradictory propositions;
- the *interpretation* of acceptance in the model is belief in the proposition (where this can be a matter of degree).[21]

ECHO is a powerful model for capturing our scientific reasoning. Again, though, it can become still more useful when integrated into a system where the explanatory coherence serves belief-desire coherence. As I have suggested, BDC can facilitate the link between beliefs activated through ECHO and the facilitation relations needed by DECO. Sometimes means-end beliefs need to be revised ("oh, it *is* possible for machines to fly!"), and sometimes it is the goals that need to be revised instead ("oh, I *can't* be a pirate!"). BDC can serve as the arbiter of such matters by testing to see which mechanisms for determining revisions works better.

Also, as we saw in section 1.2.5, ECHO must leave as primitive certain key parameters to compute explanatory coherence problems: parameters for "simplicity impact", "analogy impact", "skepticism", "data excitation", and "tolerance" (see p16 for more details). ECHO has these parameters set somewhat arbitrarily, but their *correct* value is of great

[21] Thagard (1992, 2000).

interest to epistemologists. ECHO is not on its own capable of learning better values. Adjusting them for the better requires an internal measure of error; these parameters should all be fine-tuned toward some kind of cognitive *success*. The degree of belief-desire coherence, of course, could provide exactly such an internal measure of error.

The ideal degree of tolerance, in particular, is related to another problem for ECHO that BDC might be able to help solve. It is somewhat frustrating that ECHO must take the explanatory and contradictory constraints as primitive. Input to ECHO consists of lines like this:

```
(EXPLAIN (H1) E1)
(EXPLAIN (H2) E1)
(CONTRADICT H1 H2)
```
[22]

(These can be adjusted with *degrees* of explanation, giving the weight of the constraint.) Just like the formation of the facilitation relations in DECO, forming these constraints themselves requires a great deal of cognitive work that remains mysterious. What determines which two nodes stand in an explanatory relation, or in a contradictory relation? Why should the weights between cognitive nodes be formed in that way, rather than some other? Presumably the answer comes from the intelligence embedded in the design of the network. Constraints form that way because it works for them to do so; they helped meet some higher standard of cognitive success. Either these constraints were *learned* in the service of an internal standard such as belief-desire coherence, or else they came pre-wired with the design of the system toward achieving an external function.[23]

[22] Chosen from Thagard (1992) p75.

[23] Thagard also has coherence models for deductive, analogical, perceptual, conceptual, and emotional reasoning; again see Thagard (2000) for an overview. All these can similarly be subsumed by BDC, I believe, except for the emotional coherence—on this, see section 3.2.4.

3.2.2 Explanations, contradictions, and BDC

So constraints consisting of explanatory or contradictory relations could simply be hardwired into an intelligent system, and our own human brains *might* be an example of such. On the other hand, it is also plausible that explanatory and even contradictory relations are learned. How might these relations, primitive to ECHO, be derived from a cognitive system's more basic conations and cognitions through seeking belief-desire coherence?

The formation of explanatory constraints is the more complicated example, I think, but not completely mysterious. For example, suppose it is correct that *causal* relations are basic to explanations. Learning causal relations—forming constraints that track causality—is a significantly simpler problem (still tough, of course). It may even be that something like Hume's associationist picture is right; perhaps firing of some nodes closely conjoined with the firing of other nodes reinforces the connection between nodes. This sounds very like a chestnut weight adjustment technique in artificial neural networks—the "Hebbian rule"—and it is known that at least some real neurons learn this way. Of course causality is unlikely to be quite that simple, and it is also unlikely to be easy to generalize causal constraints into explanatory ones. Still, it seems an avenue for fruitful investigation.[24]

Explaining the learned formation of negative constraints between contradictory cognitions should be easier, but it is plenty complicated. Why might a system looking to maximize belief-desire coherence form these further constraints? To answer this question we must figure out what it means, in terms of the BDC model, to hold contradictory cognitions in the first place. Perhaps a basic contradiction in this model would be to both accept and reject a cognitive node that p. Of course on the assumption that these nodes

[24]Thagard briefly suggests that explanatory relations may be causal at base in Thagard (2000) pp68–69; his own longer defense is Thagard (1999) chapter 7.

have real physical realization of some sort, no one node can have a net activation and a net inhibition at the same time.[25] So maybe instead there could be two cognitive nodes that p, one accepted and the other rejected.

But first, such a situation might not make conceptual sense—can a cognitive system have two token cognitions that p? Even if possible, such a circumstance would be guaranteed to cause incoherence among beliefs and desires if there is any conation to the effect that p around, rejected or accepted. It may furthermore be that for any cognition there is always *some* corresponding conative attitude toward it. (My proposal that cognitions arise in pairs with conations from learning mechanisms surely suggests so.) If so there will always be incoherence resulting from holding this kind of contradiction, and thus contradictions will always be unjustified in the current sense. If not—if sometimes there is no correlated conation either way, and not even the potential for one—then it is hard to see any possible harm to that specific contradictory belief state, and so it is hard to see the normative force behind any admonition against it. And anyway, if it is indeed possible to have two cognitions that p with opposite activations, *and* no potential conative attitude toward p, then the circumstance is surely rare enough to justify a general constraint-forming rule to the effect that any two cognitions that p should have roughly the same activation level.

Suppose, then, that either conceptually or because of pressures like those above, there can only be one cognitive attitude toward p at a time. Another way to understand contradictory beliefs in terms of the model is to have one cognitive node for p that is accepted at some times and rejected at others, back and forth. This approach better captures the phenomena, since people accused of contradictory beliefs do not actually accept and reject the same proposition in the same breath. Instead, they assert p while concentrating on one

[25] The use of terms like 'node' and 'activation' emphasizes the connectionist architecture I take for granted in the background, and some of what follows depends on such realizations. In symbolic systems, a contradiction would probably be realized as having symbols for both p and $\sim p$ in the "belief box". Some of what follows applies just as well to this case.

set of issues, and then reject p while concentrating on another set. In terms of the model, the activation level of p may vary to satisfy different *local* belief-desire coherencies. Such conflict between local coherencies automatically suggests a lack of *global* coherence, and so according to the theory, the cognitive system is relatively unjustified. Settling on a globally satisfactory activation level for p would represent both coming to a state of more coherence (and thus a more justified epistemic state) and at the same time represent giving up the contradictory belief. That is, holding contradictions in this sense is unjustified according to the theory of belief-desire coherence.

Still another way to understand contradictory beliefs in this model is to accept (or reject) simultaneously different propositions that are together inconsistent; the simplest case is to accept the cognition p *and* the cognition $\sim p$. Perhaps the nature of negation is such that to accept $\sim p$ just *is* to reject cognitive node p. Even if so, one need merely think of slightly more complicated examples of contradictory propositions, such as the set $\{p, p \rightarrow q, \sim q\}$. In other words, we would like to know how belief-desire coherence might be a force toward the deductive consistency of beliefs. This is the trickiest version of the question, and I can only gesture at an answer.[26]

At bottom is the difficulty of explaining how constraints could form between nodes with different contents. To make the right connections the system would somehow have to "know" which propositions are semantically relevant to which others. Thus a satisfying answer to this problem would require a good, naturalistic story about what makes for mental content in the first place—naturalistic because no other kind is presumably available

[26] Yes, Thagard does have something to say in the area too; in what he claims is the tradition of Bertrand Russell, he suggests that deductive relations are also a matter of coherence. (See Thagard (2000) section 3.3; he cites passages from Russell (1907) as evidence of Russell's view.) Deductive coherence problems have propositions for elements, deductive and contradictory relations as positive and negative constraints, and initial acceptance of propositions with "intuitive priority". It is often pointed out that deductions are really just the discoveries of inconsistent sets—after that, there is still the matter of which propositions to accept and reject. Thagard's deductive coherence seems aimed at that point; it still leaves deductive and contradictory relations as primitive, and the formation of *those* is the current interest.

to the unintelligent constraint-forming mechanism. For example, if contents of complex propositional attitudes *consist* of their place in a network of such constraints—as the conceptual role semanticist might have it—then this job is easier. There is an inhibitory weight between the cognition that p and the cognition that $\sim p$ because if there weren't (or weren't causally disposed to be, or functionally supposed to be, or . . .), they would not properly be said to have the content of p and $\sim p$ in the first place. Similarly acceptance of a cognition like $p \rightarrow q$ may simply consist in a positive, asymmetric weight transferring activation from the node for p to the node for q.[27]

Though some gesture towards an answer, this still leaves open the question: why did a constraint form between these two nodes at all? The at-bottom answer to this at-bottom question, I think, can only be: the external feedback that designed the creature demanded it be so, in order to perform its external functions. Not *all* of our cognitive dispositions can be learned from scratch. For an intelligent creature to form basic constraints between different propositions—or for it to have basic propositional content at all—must at some point have been a good way for it to turn the input of external stimulus into the output of rule-wise functionally effective action. (Of course the designer, whether an engineer or Mother Nature, can only guess at what would be effective action given any inputs, and dispose the agent toward it; that is how primitive conations arose in us.) *If* it is a basic fact about the human brain's wiring that we are disposed to accept a proposition of form q when we accept others of form p and $p \rightarrow q$, then it is because we are designed to do so. And we are designed to do so because it helped get us what we primitively *wanted*. That is, evolution gave us that primitive sub-function presumably as part of its designing us to achieve our fixed external functions.

[27] Of course if the physical realization were that simple, it would not capture the logic of the indicative conditional; the contrapositive $\sim q \rightarrow \sim p$ would have to be learned separately somehow. But actually that is not so implausible; people reason a lot more naturally with *modus ponens* than with *modus tollens*, for example (Pollock and Cruz (1999) p156 cites studies to this effect).

3.2.3 Folk psychology and BDC

So much for technical aspects of BDC, and its flirtatious overtures toward cognitive science. I would like to turn now toward implications BDC has for more traditionally philosophical problems. One such problem is the place of folk psychology in a theory of the mind. What are "beliefs" and "desires"? Throughout I have been using those terms loosely where I mean "cognition" and "conation" instead; it is time, in this section, to be a bit more careful. While we're at it, what are all these other "propositional attitudes", like hopes and supposings? David Velleman has some fruitful suggestions in Velleman (2000), on which I will build with suggestions of my own.

Velleman starts with a basic distinction in mental states between cognitions and conations, which of course is fine with BDC. Then beliefs, he claims, are the "realistic" cognitions, and desires the "realistic" conations. A wish, such as to be a 17^{th} century pirate, is an *un*realistic conation. It is not clear, however, how to spell out what makes a thought more or less "realistic". Velleman suggests

> Among the thoughts that we are disposed to make true—that is, among our conations—we delimit a subset whose members we are disposed to revise, discard, or at least reclassify if we cannot actually make them true. These reality-tested conations are our desires, which interact with one another in relative isolation from our mere hopes and wishes. Similarly, among the thoughts for which we have a disposition to behave as would be desirable if they were true—that is, among our cognitions—we delimit a subset whose members we are disposed to revise, discard, or at least reclassify if they aren't actually true. These reality-tested cognitions are our beliefs, which interact with one another in relative isolation from our mere imaginings. Setting our desires and beliefs

> apart from our wishes and imaginings is the first step toward mastering the distinction between fact and fiction.[28]

But the internal pragmatist cannot straightforwardly take on standards such as whether we can "actually" make desires true, or whether beliefs are "actually" true. Conations and cognitions can only be classified internally by which ones it *appears* can be made true or are true, and a further story is needed about what it is for a belief to appear true or a desire to appear true-able. The question then is: how do we *learn* to master this distinction between fact and fiction? The hypothesis from BDC is that to call a conation or cognition "realistic" is to say that, when compared to nearby content, it can fruitfully be added to other "realistic" thoughts to help minimize overall belief-desire tension. In a connectionist implementation of BDC, for example, a "realistic" thought is one highly activated after sufficient iterations of the coherence computation.

So a cognition strongly accepted in a BDC calculation we will interpret as a "belief", and a strongly accepted conation we will interpret as a "desire". These notions will have vague boundaries, of course—I think a feature of this characterization rather than a bug. For that matter the line between cognition and conation will be somewhat vague, on the sub-propositional model of BDC, and again I take this to be an advantage for any construal of folk psychology, which is itself surely vague.

The variety of unrealistic thoughts will have somewhat more complicated interpretations. Weakly accepted cognitions are any of the variety of cognitive propositional attitudes other than belief, such as hypothesizing that p or fantasizing that p. The BDC interpretation of folk psychology does not distinguish among these non-belief cognitive propositional attitudes based on their degrees of acceptance, but rather based on their nearby constraints. For example, suppose a cognitive node that is normally strongly re-

[28] Velleman (2000) p263.

jected temporarily reaches a low acceptance activation because, in a bout of incoherence, nearby conations strongly excite it, overruling the nearby strongly inhibiting cognitions. Then I would say that cognition counts as a fantasy. The fantasy is an unjustified cognition, resulting as it does in temporary global incoherence for the sake of some local coherence. If on the other hand some cognitive node reaches a low and stable activation as a result of balancing excitatory and inhibitory cognitions, with little or no effect from nearby conations, then that node would represent a tentative hypothesis.

Similarly on the conation side of things, I would interpret those with low acceptances as wishes or hopes rather than robust desires. Perhaps a node that reaches low activation as a result of temporary incoherence from strong nearby excitatory conations counts as a wish, while a node that achieves stable low activation from weaker nearby conations and absent or weak nearby cognitions counts as a hope. Naturally a cognitive system can reach such a state of incoherence that what would normally be a fantasy or a wish could become activated enough to count as a belief or a desire. Presumably that is just the kind of thing that happens in schizophrenics, for example. Since those beliefs and desires result from a wildly incoherent system, they are unjustified.

You might object that a hope seems really more like a cognitive state, or that a fantasy seems really like a conative one—and I agree that hopes, wishes, fantasies and the like are hard cases for the cognitive-conative distinction. Partly, I am willing to stipulate a use of 'hope' that may be usefully revisionary. But I also think the belief-desire coherence model can explain the difficulty in the distinction. In the cases of unjustified, fanciful thoughts such as fantasies and wishes, an eruption of node activation on one side of the conative / cognitive boundary will automatically tend to activate the correlating thoughts on the other side, since it is basic to belief-desire coherence that there are always excitatory constraints between the cognition that p and the conation that p. As a result, fantasies and wishes tend

to happen together, and so are hard to tell apart. In the case of hopes, it certainly seems like a wanting of some sort; their cognitive overtones are because, I think, to be a hope at all there needs to be also an associated further hypothesis or fantasy that p might in fact be the case. Maybe the best way to describe the situation is that states such as hopes, fantasies and wishes are really constituted by clusters of thoughts related in certain ways.

Finally, what about near-indifferent desires? In this model, beliefs and desires are only the strongly accepted cognitions and conations. For a cognition, it is natural to interpret a weakly accepted node as a hypothesis rather than an outright belief (where the node is stably accepted as a result of conflicting nearby cognitions). On the other hand, a weakly accepted but stable conation—resulting from conflicting nearby conations—still looks like a genuine desire. I think this apparent difference is milder than it may seem, though. A mild desire, like the "mild belief" of a hypothesis, I think is one that is only tentatively desired—*maybe* I want it to be that p; the conative equivalent of "evidence" seems on balance to suggest I would prefer that state, but it is unclear. We do not happen to have a word for the conative equivalent of a hypothesis, though perhaps 'hypotelis' would do. At any rate the states seem similar enough to justify the cognitive / conative symmetry.

3.2.4 Emotions and BDC

Naturally there are mental states other than propositional attitudes in folk psychology, perhaps most notably emotional states. These seem to be getting more attention from philosophers lately, due in part to empirical work by the likes of Antonio Damasio.[29] They have also been getting more attention from cognitive scientists lately.[30] What might BDC predict about the nature of emotional states?

[29] See Damasio (1994).

[30] See for example Jonathan Gratch's work on emotional modeling, at `http://www.ict.usc.edu/~gratch`. Meant initially as a superficial feature of the virtual agent, emotions slowly became more and more fundamental to their planning procedures. See also the recent overview Dalgleish (2003).

I have already suggested that high belief-desire coherence manifests itself as what we might call subjective happiness. Perhaps, in fact, subjective happiness just *is* belief-desire coherence, or the positive feedback that results from that coherence.[31] Similarly, subjective unhappiness plausibly results from a great deal of belief-desire incoherence. Nothing is more miserable than having fundamental desires that are not being met.

Of course there are many types of happy and unhappy emotions. (We tend to have more terms for the latter.) Each of these, it seems to me, can be a specific type of incoherence in belief and desire. For example the emotion *fear* is perhaps the reporting of an incoherence between desires to do with safety on the one hand, and beliefs about impending danger on the other. *Guilt* might be the report of an incoherence between desires to be good and beliefs that one is bad. And so on. The result is much like the *cognitive appraisal theory* of emotions, according to which emotions are (or are the result of) comparisons between cognitive and conative appraisals of the situation.[32] But according to the BDC version, these "cognitive appraisals" do not have to occur at the *propositional* level. Emotions can result from the cognitive proprioceptions of low-level bodily processes as well, as Damasio would have it. And the BDC account can also explain in part why emotions are so central to decision-making. According to BDC, emotions amount to feedback about how the agent is doing. Without such feedback, it is understandable why an agent like Phineas Gage had a hard time attaining his desires.[33]

As we might expect, Thagard has his own theory of emotions, and his model of "emotional coherence" deserves a sidenote. Thagard's take on emotional coherence is different in kind from the others; it requires adding separate "valence" values for nodes in addi-

[31] Or perhaps even the qualitative experience that results from that feedback, though I doubt that is the best way to talk.

[32] See Ellsworth and Scherer (2003) for a recent overview.

[33] As pure speculation on the matter, note that the mysterious dopamine is implicated not only in learning and reinforcement, as mentioned earlier, but also in positive affect. Also see footnote 14 on the role of the anterior cingulate cortex in emotion and learning (p74).

tion to their activation levels, and separate weights between nodes for flow of emotional valence.[34] As I have suggested, my guess is that emotions are not matters for coherence, but are themselves manifestations of internal monitoring of coherence and incoherence. Thagard seemed at one time to agree with such a view of emotions; for example he and Millgram say:

> We conjecture that subjective well-being is as much a matter of coherence among one's actions and goals as it is a matter of accomplishing goals.[35]

Of course I think subjective well-being arises from perception of goal accomplishment, in the form of belief-desire coherence, and not from the mere formation of a coherent plan. But the point here is that it looks like Thagard also is tempted to think that emotions arise from kinds of coherence, rather than being a matter for coherence themselves.

And Thagard does agree that at least some emotions correlate closely with levels of coherence:

> The usually pleasant feeling that something makes sense involves an overall assessment of coherence, in contrast to the confusion and anxiety that often accompany incoherence. I call these *metacoherence* emotions, because they require an overall assessment of how much coherence is being achieved.[36]

He also gives surprise as an example of a metacoherence emotion. BDC proposes that *all* emotions are "metacoherence emotions". This is more plausible when the coherence or incoherence is between beliefs and desires, rather than the various kinds of cognitive or conative coherence Thagard details.

[34] See Thagard (2000) chapter 6.

[35] Thagard and Millgram (1995) p451.

[36] Thagard (2000) p193. BDC would, for example, have the emotion *confusion* be a reporting of incoherence between desires to understand and beliefs that one doesn't. And so on.

3.2.5 Ethics and BDC

Though it is venturing somewhat far afield, there are three points of intersection between BDC and ethics that warrant a little immediate attention. First, on the defensive side, I want to clear BDC of sociobiological implications it may seem to have. Second, I would like to suggest two possible uses for BDC in ethics: as a possible foundation for utilitarian axiology, and as a starting place for thinking about the moral considerations due to robots.

First, on sociobiology: because BDC supposes that primitive conations in humans arose from evolutionary design, and that those serve as a foundation for all rational desires, it may seem to endorse an evolutionary approach to ethics. According to such a view, what is right is roughly what is evolutionarily successful. Murder is wrong, for example, because murderous communities are less efficient at propagating gene copies. But sociobiological approaches to ethics are naive. First, to get intuitive results they commonly presuppose mechanisms of group selection that are unlikely to exist. More importantly, it seems at least a conceptual possibility that controlling population growth, for example, might in some cases be the right thing to do. If BDC has sociobiological ethics as an implication, so much the worse for BDC.

Admittedly, default conations in humans are likely to be the ones adaptive in the old evolutionary environment, and it is easy to imagine that those are often selfish and unethical. A system aspiring to belief-desire coherence needs to take these stubborn fixed conations seriously—otherwise thinkful wishing would be the best reasoning method. But as we have seen, to take default desires seriously is not to endorse each one. Conflicting thoughts (both cognitive and conative) can inhibit conations until they become rejected, or at any rate accepted only at the level of wishing. More importantly, it seems likely that altruistic desires can be rational. There *might* be default altruistic desires in humans,

for example, if enough undiscriminating phenotypic altruism sufficiently benefited kin in the evolutionary environment. There need not be mechanisms of group selection to make sense of this; eusocial creatures such as ants serve as prime examples. But even if all fixed, default conations are selfish—or anyway genetically selfish—it may still be rational to develop derivative altruistic desires from the balance of them, especially in animals as social as humans.

According to the BDC picture, determining which conations are rational—or in the terms of belief-desire coherence, determining which conations cohere enough with other thoughts to be called a "desire"—requires cognitive effort. Richard Brandt thought rational desires could be determined through a process of "cognitive psychotherapy". Belief-desire coherence could perhaps put a more definite gloss on the nature of this process.[37] An independently grounded notion of rational desires, in turn, could provide some foothold for a utilitarian axiology. Perhaps it is only good to satisfy the rational desires, rather than any old desire. (Here I mean *actual*, not apparent, desire satisfaction.) Sadistic pleasure, for example, might no longer be a factor to consider in such a utilitarian calculation.

But what about intelligent creatures with fundamentally different aims, such as the smart robots a future could bring us? They presumably would not typically have fixed conations like "get protein". It is likely they would not always have fixed conations like gaining romantic affection, either. (After all, they need not have the innate drive to reproduce we have, and even if they did, their reproduction would not always depend on another creature.) Indeed depending on their basic aims, perhaps sadism, or complete servitude, could be rational desires for intelligent creatures. And here BDC leads me to a view on robot ethics that surprises me, and even disturbs me a little, but that I think may be right.

[37] See Brandt (1979). A related note: on p197 of his (2000) book, Thagard cites research showing coherence problems have had some success at modeling the psychological theory of cognitive dissonance. He proposes his own view of emotional coherence would do a still better job; naturally I have reason to think my own would do better still.

Yes, complete servitude could be a rational desire for a robot; it could be an activated default conation, or result derivatively from a balance of them, in a coherent state. And suppose that what makes an action wrong is an on-balance frustration of creatures' rational aims. It could then be ethical to have intelligent, robotic slaves; indeed, it would be cruel to "free" them. Unlike their 19^{th}-century human prototypes, or our modern-day animal servants, a creature actually designed to serve a human family properly *wants* to do that serving, in the only sense of 'want' I can figure. To suppose it "wants" to be free and live its own life instead is anthropomorphizing. Of course any normal *human* in that situation would want that, because evolution has given us fixed conations that slavery would only violate. But a robot slave would actually, we can suppose, be *happier* in fulfilling the goals for which it was designed. (We should of course be careful here, though; rationalization has proven too powerful a force before.)

If this is hard to swallow, it might help to think of dogs. Domestication may be part of their evolutionary design, and they similarly might actually be happier pleasing their human masters than set "free" to negotiate the wild on their own. Of course a crucial difference is that dogs are not as intelligent as humans. But if as we have supposed intelligence is about adaptability in reaching basic goals, then I cannot see how the level of intelligence would on its own affect whether servitude were ethical. If dogs do fundamentally, genuinely, rationally want to please their owners, it is just as ethical to let smart dogs do it as dumb ones. In fact it might be more so, just as it is (probably) worse to kill a human than a cow. More intelligent creatures have "more" interests, or more complex interests, in a way that seems to weigh the achievement of those interests more heavily.

Naturally most intelligent creatures will still have desires like self-preservation as importantly instrumental to the achievement of their basic aims. As such, it would be unethical to kill them, for example—that would presumably involve automatic frustration of

their rational desires. But what if, on the other hand, the robots fundamentally want to kill *us*? Since much AI funding comes from the military, we are unfortunately likely to see smart killer machines before smart serving machines. On the BDC proposal, killing, too, could be a rational desire for a creature—say an intelligent, mechanical suicide bomber. (Note this creature would not even have a rational desire for self-preservation.) But of course on the picture of ethics where on-balance frustration of all rational aims is wrong, the aims of the creatures killed would still outweigh the killer's aim.

3.2.6 Conclusion

There are lots of reasons to hope that a theory like BDC is correct. It is a cognitively realistic, internally available, and normatively appealing approach to cognition. If viable it could serve as a standard for derivative methods of inference such as logic, abduction, or practical reasoning. By allowing for degrees of belief and providing a method for choosing goals, it could supplement utility theory. It could provide a better model of what goes wrong in cases like *akrasia* and poor performance on reasoning tasks. Its hypothesis about internal feedback in the human brain could supplement and spur useful research in neurobiology and artificial intelligence. It could give ethical notions a firmer foothold. It could perhaps even provide insight into practices of clinical psychology, helping people lead mentally more healthy lives. More generally, a pragmatic approach to reasoning like that of BDC could provide a standard for philosophical debates that are difficult to ground, such as intuitions regarding counterfactuals and ontology.

All right, I admit: all that may be just a bit too much. After all, belief-desire coherence, so far, is just a *hypothesis* about the best way to describe learning in cognitive systems like us. But if research along these lines can help realize any of this potential, then it is a hypothesis worth pursuing further.

CHAPTER IV

TRUTH AND INTERNAL EPISTEMOLOGY

Of all the epistemic controversies discussed thus far—internal *vs.* external, coherence *vs.* foundational, natural *vs.* non-natural—probably the most controversial stand I have taken is to side with *pragmatic* epistemology against a traditional *alethic* standard. In the previous three chapters I have tried to provide strong positive motivation for the position. This last chapter is for those who have had "truth" buzzing in their ears that whole time. The goal of this chapter is to make such readers at least more comfortable with pragmatic epistemology.

My approach to this goal will be two-pronged. On the one hand, I will argue that truth cannot serve as an internally available standard for thought, and on the other hand, I will argue that truth would not be an internal standard with the right kind of normativity for thought. That is, in specific and restricted senses, I doubt both that our thinking *can* aim for truth and that it *should* aim for truth.

4.1 Can we aim at truth?

The title of this section is misleading. Of course, a creature *can* aim for truth—in the sense that a creature is likely to have mechanisms with the function to provide accurate information about the world it inhabits. The question instead is whether a creature can

have a feedback mechanism that adjusts its own thinking processes toward an internally represented goal of truth. Put another way: can there be an internal measure of error from the truth that will guide a creature in the way internal epistemology demands?

Some of the reasoning to follow depends specifically on BDC's "learning" version of internal epistemology, but the majority is of a more ecumenical nature. And many of the arguments apply to any internal alethic epistemology, but I also make trouble where I can for specific proposals.

4.1.1 A warm-up argument

Let me first try a short argument that goes some way toward showing why any broad internal epistemic project cannot be an alethic one. This argument is a warm-up; it does not yet rely on anything specific to BDC, and it has problems that we will explore in the next section.

First, we need a more formal construal of internal and external epistemology. Take the "demonic" and "supervenience" intuitions from section 1.2.2 as basic to internal epistemologies. They are only concerned with internal psychological conditions of the thinker, while external epistemologies make reference to the thinker's external environment. Let's put this more formally by saying that an internal normative epistemology is a partial ordering of all possible total psychological states s, while an external normative epistemology is a partial ordering of all possible *pairs* $\langle s, w \rangle$ of psychological states and circumstances. These "circumstances" are probably completely specified in some way—maybe centered possible worlds. The ordering in both cases will be partial since only some mental states are relevantly compared—for example, those with the same perceptual beliefs, say (whatever those are). An important goal of either kind of normative epistemologist is to make useful generalizations about these orderings.

Second, we need a more formal construal of an alethic approach to normative epistemology. Roughly, alethic epistemologies are committed to saying that one dimension of having a better thought is for the thought to be *true*. But this is not quite enough, since we know that 'true' gets used in lots of ways. If a pragmatist like William James says that we should aim for truth, and then adds that the true is just the expedient in our way of thinking, then we will not count his epistemology as alethic! So strictly speaking an alethic epistemology seems to rely on the good old-fashioned *correspondence* theory of truth, according to which truth is a *relation* $\langle t, w\rangle$ between a thought and a world.[1] (I assume, perhaps erroneously, that an alethic epistemology with a *deflationary* theory of truth will either accept this relational construal once talk about the "world" is properly hedged, or else will be so deflationary as not to be of concern to a pragmatic epistemologist.)

Perhaps you see where I am going here. Suppose we have an alethic normative epistemology. To be properly alethic, it will generally prefer s to s' when s has more true thoughts, and in the general case, more *empirical* true thoughts. That is, a thinker's psychological state is generally better according to an alethic epistemology when it more accurately reflects the state of affairs w in which the thinker is thinking. Psychological state s may have more of these truths than s' in w, and fewer in w'. Thus, naturally, we can in general only say whether s is preferred to s' on alethic grounds relative to circumstance w. That is, an alethic epistemology can in general only make claims of the form $\langle s', w\rangle \leq \langle s, w\rangle$. That is to say, by our definitions: if a general epistemology is alethic, it is external. So if it is not external, it is not alethic. So we conclude that no general internal epistemology can be alethic, which is what we wanted to show.

Of course we do not really need this pseudo-math talk. The point is simple: to judge psychological states based on how well they reflect the world, we need to be able to talk

[1] For example, if the proposition the thought expresses is understood as a set of possible worlds, then the relation is that the thought's proposition should *include* that world as a member.

about how the world is.

Now you might want to say that if internal epistemologies cannot avail themselves of truth, then so much the worse for them. But remember what motivates internal epistemology in the first place: the need for advice that I can use. Suppose I am figuring something out (intuitively speaking), and it comes down to a choice between psychological states s and s'. The external epistemologist can in general only tell me that I should prefer s in w and s' in w'. In some cases this may be useful information, if I already have a good guess about my circumstances, and never noticed its implications for the decision between s and s'. It is times like these that external epistemology is helpful.

But other times, I am in no better shape than I was before, since I do not know what circumstances I'm in (suppose the choice between s and s' would be easy if I did!). The internal epistemologist, though, can give me a standard for choosing independent of circumstances. It may be wrong from an alethic point of view—it may prefer s even though I am in w'—but, intuitively, it is the best a thinker can do. Internal epistemology comes into its own when we despair of getting independent grip from the inside on how the world is, and so need to determine what would make a better psychological state *however* the external world may be.

4.1.2 BDC, *a priori* beliefs, and the truth aim

There are at least two major weaknesses in the warm-up argument that leave room for an internal alethic epistemology.[2] First, it concludes that no *general* internal epistemology can be alethic. But perhaps the alethic epistemologist is willing to settle for less. There could be an internal epistemology that concerned itself with only what I would call *a priori* truths, if there are any such. That is, it could be that some thoughts t are such that the

[2] I am particularly indebted to Eric Lormand for this section.

truth relation $\langle t, w \rangle$ holds for *any* w.[3] Such an epistemology would not judge psychological states based on how well they reflect the world, but on how well they *could* reflect any world, and so should count as alethic nonetheless. For example, perhaps s' contains an extra proposition that contradicts something already in s. We may then have internal reason to think that we should always prefer s on alethic grounds. In fact logical truths in general seem amenable to internal, alethic justification independently of the thinker's world. Perhaps similarly we will always have alethic reason to prefer s if it assigns degrees of belief that adhere more closely to the axioms of probability.[4] And a psychological state s containing true beliefs about s itself will be better (all else being equal) than a similar one with fewer such true beliefs, no matter the world in which s finds itself. So for all the warm-up argument shows, an internal epistemology could provide purely alethic justification to beliefs about the internal world of the thinker. (These are included in the above sense of "*a priori*".) So even granting the argument of section 4.1.1 against internal alethic justification of empirical beliefs, the logical and experiential beliefs still available for internal alethic justification would be enough for a pretty robust epistemology. The result would be a "partial" alethic view that strives for truth-linked justification when possible, but settles reluctantly for some lower-grade justification (like pragmatic benefit) when it comes to the empirical beliefs.

A second difficulty with the warm-up argument is that there could be truth-linked reasons to prefer one psychological state over another even if the preferred one turns out to contain fewer truths. For example there may be *alethic* reason to condemn wishful think-

[3] "Those are *necessary* truths, not *a priori* ones" you might object. No. I take it that *thoughts* have "narrow" content, so the proposition thought t expresses must be filled out by context; you must first determine the content by evaluating its indexicals in the (centered) world w. *Then* you determine whether w is included in that proposition to see if the truth relation holds. In other words, in this formal (but somewhat simplistic) model, I think of the *a priori* propositions as those true in the "diagonal" of a two-dimensional modal semantics, as proposed in *e.g.* Davies and Humberstone (1980).

[4] See Joyce (1998).

ing even if, by pure accident, it resulted in more true beliefs than apparently responsible methods. It is difficult to spell out just *how* a psychological state with more falsities could be preferred on truth-based grounds, as we will see, but there may be room for such principles. Presumably, these principles could not depend on the world of the thinker, on pain of having no internally available reason to think *them* true. So the idea is that perhaps among the *a priori* beliefs are enough resources to start aiming for alethic justification when it comes to the empirical beliefs, too. For example, experiential beliefs with internal alethic justification could be used as testing material for methods of generating beliefs; it may be *a priori* that methods able to generate these beliefs (the ones we have independent, internal reason to think true) also will tend to generate true beliefs for which we have no such independent, alethic justification. Or it could be *a priori* that internally testable features of beliefs, such as simplicity, contribute to necessary components for finding true beliefs. (Indeed, we will see such a view in section 4.1.6.)

In the following sections I argue on more general grounds against attempts to build a full-blown epistemology out of the *a priori*. But first in this section I would like to give reasons specific to BDC for thinking that any such project faces serious conceptual difficulties. I believe this argument makes trouble even for internal alethic justification of the *a priori*.

In overview, this argument from BDC will rely on the content of the truth aim. Presumably the truth aim looks, in more detail, something like this: have your cognitions accurately represent the way the world is. Or, less contentiously perhaps: believe p if and only if p.[5] But in the hypothesized sense of "conation" from chapter I, the nature of this aim is problematic. It is one thing for a creature to have accurate representation as a *function*. My perceptual capacities, lower and higher, have as a fixed function something like

[5] As Lormand (2001) would have it. Lormand emphasized this is an *ideal* truth aim; most thinkers will settle for far fewer than all truths, though still holding this as an ideal.

accurate representation of the world around me. And creatures unable to learn at all have similar functions to represent the way things are, as when bees represent to their sisters where the flowers are. But on the BDC hypothesis, conations are internal representations that arise from a learning process, which in turn is understood as a feedback mechanism for adjusting a creature's mental computation. On this model conations must come with cognitions representing how the creature is doing with respect to this goal, together forming a measure of error that can be fed back to the system. But how can we get such a measure of error for the truth-aim? Put another way: what is the cognitive side of this pair, the side that represents how the thinker is doing when it comes to the truth-aim? Or, in effect: how can we learn to get truth through a process of trial-and-error? The error, on any mind-independent construal of truth, simply is not accessible to us.

So now for the argument itself. It has the structure of a dilemma. Consider some cognitive mechanism M that putatively forms beliefs with internal, alethic justification—for example, a mechanism for avoiding contradictions. Was M learned, or not? Suppose it was not learned; suppose instead that it came hardwired with the creature, or was the result of a mad scientist's midnight neurosurgery, or developed from a serious blow to the head. Then according to the BDC construal, M did not result from an internal epistemic norm; it was not formed by internal guidance toward some goal. In particular, then, it was not formed by internal feedback toward a *truth* goal. Of course a mechanism like contradiction avoidance will in fact contribute toward having only true beliefs, and contributions toward true beliefs may even explain why M is there. But these reasons will not have to do, it seems, with the creature providing itself epistemic guidance through internal adjusting of its own cognitive mechanisms, and so the justification of the mechanism cannot point to any internal aims of the thinker. So in this case we could have alethic, but not internal, justification. (Section 4.3.1 will discuss this kind of justification a bit more.)

Suppose, on the other hand, that M *was* learned. Could it have been learned through feedback from an internal aim at truth? We suppose the internal truth aim takes a form like "believe p just in case p." According to BDC this internal aim is itself a mechanism, call it T, which provides feedback to its sub-mechanisms, encouraging some and discouraging others. So according to BDC, to suppose that M has internal, alethic justification is to suppose that M formed through positive feedback from T. But how could this work? The thinker hazards M as a new mechanism. T tests M for whether it helps bring about the goal of believing that p when and only when p. In effect T must compare the old belief set B, formed without M's help, and the new belief set B' produced at least in part by M. And T compares these belief sets on the basis of which one better fulfills the goal of believing according to the way things are. But of course the thinker cannot directly compare its belief set to the way things are; it can only compare its belief set to its *representation* of the way things are. But that representation is just B! And B will always compare better to B than to anything else. So any new mechanism will lose out on T feedback. T would be, if anything, a rigidly conservative force; no new belief-forming mechanisms could be learned in its service, and any belief-forming mechanisms formed in the service of *other* aims would only be endorsed by it. On this horn of the dilemma we could have internal, but not alethic, justification.

Take the specific example of a contradiction-avoidance mechanism. It scours the belief set B and eliminates contradictions, and reports a new contradiction-free belief set B' to T for evaluation. T checks to see if B' better reflects how the world is by comparing it to its representation of the world, B. Not surprisingly, B' loses. But perhaps, you might object, the thinker already has a principle that p and $\sim p$ cannot both be the case. Wouldn't it *then* endorse the new mechanism, seeing that it rids B of contradictions that could not accurately reflect how things are? But it seems that for the thinker to have

such a principle is for the thinker already to have, in effect, the contradiction-avoidance mechanism. The thinker *already* prefers belief sets free of contradictions on other grounds, and T simply rubber-stamps it. Or take the example of a new mechanism that forms beliefs from experiences. If the thinker did not already have representations of such experiences, it seems T cannot endorse the new mechanism as better reflecting how things are.[6]

Perhaps, though, there can be independent and internally available reason for thinking that some mechanism would bring about true beliefs. Instead of forming a mechanism that actually scours beliefs for contradictions, a thinker could just somehow come to believe specifically that "eliminating contradictions will contribute toward believing truth." A thinker who does not explicitly aim to believe truths would be unmoved by this new belief; only in *combination* with T would the new belief actually result in a new cognitive device for eliminating contradictions. But how does the thinker get the belief that contradictions lead to falsehoods—from what belief-forming mechanism? We have seen reason to think the mechanism could not arise under feedback from T itself. So it must have come from some other feedback mechanism, for example "believe what philosophers say." And we must suppose, from the same argument above, that *this* mechanism did not arise in the service of truth. So a new mechanism for avoiding contradiction arises, but not from T alone—the truth aim had to make essential use of a non-truth-based reason. So the justification for the new mechanism does not seem genuinely alethic.

Maybe the alethicist would be satisfied with that effective a truth aim. Then again, maybe the alethicist has some other trick that has not occurred to me. But the BDC theory gives at least reason to doubt that thinkers can aim for truth in a substantive way. And

[6]I should acknowledge that this argument focuses on belief-forming *mechanisms* and the justification they endow on beliefs. Perhaps a belief could have internal, alethic justification, but not in virtue of the mechanism that brought it about. But it is hard to see how. To internally justify a belief, it looks like the justifiers have to be involved somehow in bringing about the belief. But then it looks like there is a mechanism for taking those justifiers to a justified belief.

even if we can, BDC also gives us reason to think that the explicit aim for truth will be an instrumental one, itself formed in the service of higher, pragmatic aims of the creature. I will discuss this point in the "should we aim at truth" portion of this chapter (specifically section 4.2.2). And again, BDC does concede that a creature can aim for truth in the sense that it very likely was built with truth-producing mechanisms. But this is an external cognitive aim—one that does not provide internal guidance.

4.1.3 Evidentialism

So we have some reason to discount the prospects for an internal, alethic epistemology. Still, though I am not *alone* in supposing that internal epistemologies cannot be alethic, truth-based accounts of internal justification are hardly unpopular. The standard approach is *evidentialism*, which seeks to identify special elements of the (internally accessible) psychological state that count as evidence.[7] Broadly stated, evidentialism deems someone's belief justified just in case the evidence available to her supports the belief. The internal aim, then, is to believe according to evidence. Normally this is an alethic view because on the typical understanding of "evidence", possession of evidence for p is somehow connected positively with the truth of p; thus one can aim to believe truly by aiming to believe according to evidence. But, pragmatists are quick to point out, it is hard to identify in a principled way what beliefs count as evidential support in this sense. I will expound this problem, and show how it stems from a deeper one.

Let's start with the shipowner example from W. K. Clifford, the proto-evidentialist.[8] A shipowner considers whether his vessel is seaworthy. The vessel is old and leaky, having made the voyage several times—this last time, suppose, rather shakily. Intuitively the leaks

[7] See, for example, Feldman and Conee (1985).

[8] A modern evidentialist need not claim with Clifford that it is *always wrong* to believe in a way that does not conform with evidence; epistemic justification and the all-things-considered right thing to do might come apart. See Clifford (1879).

and previous difficulty stand as evidence that the ship would be lucky to make another run. But why don't all those successful voyages, especially the lucky ones, stand as evidence that a guardian angel is watching over the ship? Or why don't the apparent leaks stand as evidence that some competitors are trying to trick him into believing his ship is not seaworthy? Or perhaps the fact that other ships in this condition tend to sink is evidence that every ship in such a condition *except his* would be likely to sink? For that matter, why isn't the fact that he *wants* the ship to make another run evidence that it will?

One way to put the point is that several theories are consistent with any set of observations (granting we can work out a non-theory-laden notion of *observation*). Yet presumably not all such theories will be justified by the "evidence". Why are observations *evidence* only for some small subset of those theories? To pick that subset is to appeal to principles that go beyond the observations themselves. In other words, to make sense of evidence in a way that rules out the putative evidential relations above, we must make sense of how theories can be better or worse on grounds other than consistency with the data. We need "theoretical virtues" such as explanatory power, or conservativism, or simplicity. But it is problematic to show that such principles depend on the aim for *truth*. Certainly they have pragmatic value; it is easier and more efficient to have simple theories, for example. Maybe the pragmatic value of the simpler or more conservative theory is a happy, coincidental side-effect of principles chosen for purely for their alethic qualities—but many philosophers will confess this at least looks suspicious. In fact it is hard to see any reason at all for saying the simpler theory is closer to a true one.[9]

So to show that evidentialism has truth as its internal aim, we need to characterize better theories—and thus, indirectly, evidence—in a way that relies only on evidence's connection to truth. But just what is this target connection between truth and evidence?

[9]Though we will examine one possible such reason in section 4.1.6.

A first, naive pass at this problem would construe the proper connection to be a simple positive correlation—the more evidence one has for a belief, the closer to the truth it is, period.[10] But this is clearly no good. If the notion of evidence is so strong that it is sufficient for true belief, then evidentialism cannot make room for the possibility that one can have evidence for a belief that is false. But that would fly in the face of intuitions fundamental to internal epistemology, such as the "demonic" intuition from section 1.2.2. Intuitively, we could be doing a pretty good epistemic job and still have false beliefs. (Perhaps you simply do not share this intuition; further investigation of it will come a bit later, in section 4.1.5.)

All right then, we suspected that route was too simplistic. But what else is available to the evidentialist? Perhaps, say, evidence is connected to truth because an evidential state makes the belief more "probable"? But this proposal too is problematic for internal epistemologies. What does it mean to say that evidential state e makes belief p more probable? Maybe it means that 19 times out of 20, when you have e, then p. But first of all, it is not clear how such a probability gets reckoned. As long as we allow for deceitful demons, for any evidential state there will be countless cases where p fails to hold—at least, as long as we assume that evidential states are mental states. We might attempt to restrict our attention to "normal" worlds, but it would not be easy to specify these without begging the question. Suppose, though, we manage somehow to make sense of e tending to be present only when p. Still, for the proposal to be internal, not only must e be cognitively available to us, but it also must be available to us *that e is an evidential state* in this sense. Without this condition the version of "evidentialism" under consideration is basically reliabilism—a clearly external epistemic project. We wanted a notion of evidence that could provide internal feedback. To aim for the external, relational goal of truth, the

[10] If talk of approximate truth make you edgy, consider instead the even less plausible view that beliefs with (enough?) evidence just are true.

evidentialist tells me to aim internally for believing according to evidence. But which mental states are evidential in the way proposed is not available to me; I have no way to assess these objective probabilities. I might have *guesses* about what states are evidential, with associated subjective probabilities. I might be in state e, and further believe that "19 times out of 20, when I am in state e, then p." But if a subjective evaluation of e as an evidential state is all that is required for the state actually to be evidential and thus justificatory, then we have lost the connection between evidence and truth, since surely the subjective evaluation could also be wrong. Furthermore, the evaluation may need its own justification through some evidential status in order to be justified internally, leading to a worrisome regress.

4.1.4 More general problems: content

I would like to mention here a more generic worry I have about internal aims at truth. Any reasonable such attempt to aim at truth will naturally make use of the belief's content—for as Eric Lormand puts it, the content of a belief is in a necessary dependence relation with its truth or falsity.[11]

But just this simple point can cause problems for an alethic view. Content is tricky, and we are not really sure what it is. That is no objection so far, of course. But one of our best guesses about the nature of content is the "teleosemantic" line.[12] This line, I suspect, is incompatible with an alethic project. For according to teleosemantic views, the "consumer" side of a representation determines its content; the content depends on the aims of the creature representing. Consider a hungry bird. In her environment, there are tasty butterflies, and there are poison butterflies. The tasty butterflies attempt through selection pressures to mimic the poisonous ones, and the poisonous ones attempt to change

[11] Lormand (2001) section VII.

[12] See for example Millikan (1984), Sterelny (1990), and Papineau (1993).

their patterns to avoid this free-riding, and so on. In the midst of all this pastoral confusion the bird might accidentally eat a poisonous butterfly. But the bird has misrepresented this butterfly as of the tasty variety, not correctly represented it as any butterfly that looks that way. That is because what explains the bird's mental state, according to the teleosemantic line, has to do with the normal explanation for those representational facilities. The normal explanation is that it helped the bird's ancestors get lunch, because normally what's eaten is a tasty butterfly.

Now a true alethicist seeks to insulate the truth search from the influence of passions, forswearing any pragmatic influences.[13] But if teleosemantics is right, the content of beliefs is *determined* by these aims, and so to discriminate beliefs by content is in part to discriminate according to aims. The alethic quest is polluted from the ground up by pragmatic concerns.

Consider a typical evidentialist's predilection for beliefs that exhibit analogy, for example. If our bird were an evidentialist, and if she somehow had no aims to eat lunch, then perhaps our evidentialist bird could assimilate representations of the various butterflies around her into one category: butterflies that look roughly like-so, perhaps. There would be one kind instead of two, and that would exhibit greater analogy than breaking them down into further categories. But the bird's aims keep her from having such content. She must try to sort the butterflies around her, confusingly similar as they are, into *two* categories. She must because she has aims of eating only the tasty butterflies.

Similarly a human might like to assimilate all the clear splashy thirst-quenching stuff into one kind, water. But if it should turn out that his environment contains the clear splashy XYZ as well as regular H_2O, it may be that his *interests* dictate he keep these categories separate. (Other times his interests may allow him to assimilate, as in the case

[13] I have in mind a truth-seeker like Lormand's Renée, discussed in sections 4.1.6 and 4.1.7.

of jade. Section 4.2.5 will expand on this point.)

Of course there are alternatives to the teleosemantic approach to content. The evidentialist could hope instead that some "upstream", information-semantics approach to content like Jerry Fodor's is the right way to go.[14] But to the extent that teleosemantics is a promising route to determining content, the evidentialist's hopes of isolating practical aims from the epistemic project seem optimistic.

Well, perhaps the typical evidentialist is just unfortunate to be in a human body with humanly aims. And other creatures with other worldly aims will similarly be unfortunate in having any purely alethic aspirations infected at a basic level. But suppose we could build a creature with one basic aim: to gain the truth. *This* creature would escape the worries of pragmatic aspects to content, surely.

But the problem is, what would such a creature even look like? How would we go about building such a thing? Would it count, for example, if we duct-taped together various recording devices—maybe a video camera, a seismograph, a recording thermometer, and so on? Presumably not—that would not seem to be a creature, and it would not seem to have any content. Well, how then? Perhaps add to these perceptual devices dicta like simplicity and analogy and parsimony? But then, if the teleosemantic line is right, the creature could have ideal analogy and parsimony by simply not making any categorical divisions at all, representing everything as one entity. There would be only one potential belief to consider and score. Perhaps such a "representation" would even be true. But I suspect at that point we would lose interest in truth.

[14] See Fodor (1987). The upstream / downstream distinction in theories of content is from Peter Godfrey-Smith; as he puts it,

> indicator theories [like Fodor's] are 'upstream-looking' theories; they link a representation to its object by looking to the processes involved in the bringing about of the representational state. They are based upon the organism's powers of discrimination by means of perception. Teleological theories look instead to connections 'downstream' of the representational state, connections going via behaviour and its consequences (Godfrey-Smith (1998) section 2).

Besides, there would be countless different ways to "simplify" (or analogize or ...) the incoming data. For example, the creature could keep every other datum, or the first thirty of any group. Would any one of these simplifications suddenly give the creature content? If not, then why some and not others? The intuition that such a creature cannot derive content from a pure truth aim, if you share it, speaks strongly for "downstream" approaches to semantics. Perhaps the device's processed data could even have asymmetric dependencies, as Fodor's upstream theory would want it, but it is still not clear how it could genuinely represent and misrepresent.[15] And to the extent that we have reason to prefer downstream content theories, it seems to that extent we should have doubts about internal alethic aims.

4.1.5 More general problems: means to truth

Marian David has another general problem for an internal, alethic epistemology.[16] We are supposing that satisfying some internal aim—believing according to evidence, say—will be a means to satisfying an aim for truth. But in what sense can satisfying the one aim be a *means* to the other? It could be a causal sense. But

> a person can be justified in believing p even if believing p will cause her to hold a massive amount of false beliefs later on. ... More generally, it seems that being justified in believing p has nothing to do with what beliefs you are going to hold in the future. The truth-goal cannot be a diachronic goal if it is

[15] Suppose for example the device took several visual inputs per second, and was set to use the first twenty or so inputs of its existence to determine a subfield of those inputs to store, in order to "simplify" the data. The privileged subfield could be determined by the darkest patch, or the lightest, or the most varied, or the most constant, or what have you. And suppose the first twenty inputs to this machine are all the same image. Now the machine has locked in on a subfield for simplification purposes. Later the machine has a different input that happens to coincide with the initial inputs on the subfield the machine has chosen, but differs outside it. The machine will process this data just as the initial images, though from a different cause, and its processing that data that way depends asymmetrically on the initial cause. Still, does that make the "simplified" data contentful? Would it make sense to call it false, or a misrepresentation?

[16] See David (2001), especially section V. It is based in part on an argument by Stephen Maitzen.

> to play the role assigned to it in the goal-oriented approach to justification; it cannot be the goal of having beliefs that are true and not having beliefs that are false *in the long run*—if it were, the causal consequences of our beliefs would be relevant to their epistemic status. Instead, it must be a synchronic goal: It must be the goal of *now* having beliefs that are true and *now* not having beliefs that are false.[17]

The diachronic approach seems to have serious difficulty. But on the other horn of the dilemma, David says, the only apparent way justified beliefs can be synchronic means to believing truly is as a constitutive means—the way bricks and boards arranged in the right way are a synchronic means to a house. But on a constitutive account "justification collapses into truth", since "a constituent of the goal must always be a better constitutive means than a nonconstituent."[18] Any true belief, however rotten its intuitive justification, is a better synchronic means to the truth goal than its competitors. And any false belief, however golden its intuitive justification, is a worse synchronic means to the truth goal than its competitors. All and only true beliefs are justified on this view.

David calls this last the "*reductio*" argument. For a further diagnosis of the problem, it is worth examining the absurdity to which the position has been reduced: that justification and truth collapse into each other. Why is this so bad? Put another way: why is it so important to *dis*connect the two (at least a bit)? Intuitively, we think one could be doing a pretty good epistemic job and still end up with a false belief—or vice-versa. That is basically the demonic intuition of section 1.2.2. If so, then there must be an epistemic good other than truth. That good might *depend* on a truth aim in some way. But it cannot depend on it so heavily that it violates the intuition that truth and justification are separate.

[17] David (2001) p160. Presumably David has in mind diachronic causal problems like Gilbert Harman's example of believing a falsehood in order to get an encyclopedia full of truths, thereby justifying the false belief; see Harman (1999).

[18] David (2001) p161.

What is so sacred about this intuition? At root, I think, is at least a mild form of realism that almost everyone shares: the presumption that our mental states could get the world (or something extra-mental, anyway) *wrong* somehow. For a belief that p, the world could be such that p or not. There is a one-many relation between psychological states and the world. And fundamental to *this* intuition is that we often are forced to revise beliefs, even ones in which we were once very confident. We have all been burned in this way. As a result we have come to suspect that no amount of confidence is sufficient to guarantee unrevisability. We gain skeptical intuitions, and allow for the possibility of massive deception, such as that of a demon world. In explaining the need for epistemic justification, philosophers often cite a distaste for "happy-go-lucky truth"[19]—truth that is accidental and not earned. But I think this scorn is just the heroic flip side to a *fear* of getting things wrong through some failing that could have been avoided. The possibility of being wrong, the coherence of skepticism, and a modest realism are all bound together.

Again, a base intuition of internal epistemology is that in some sense the same psychological state is doing just as well in the demon world as the real world. Tied to this intuition, I have argued, is that we want epistemic norms to guide us. They should provide feedback for adjusting our cognitive mechanisms. Where there is a one-many relation between psychological states and success at an aim—as our base skeptical intuitions insist there is with truth—that aim cannot be internal to the psychological state in this sense. This is how the combination of (modest) realism and a demand for guiding norms ensures that truth and our intuitive notion of justification remain separate. Because our realist, skeptical intuitions say that success at any internally available aim could fail to achieve truth, the collapse of justification into truth is an absurdity.

[19] As David calls it, David (2001) p166.

So internal alethic epistemologies are fighting both to connect and disconnect their internal epistemic aim with truth. The internal aim must be a fallible means to the external aim, and it cannot be a diachronic or constitutive means. This is a tricky business.

Perhaps the problem David pointed out comes from conceiving of justification as a *means* to truth in such a crude way. Maybe some other conception of the connection between truth and justification will do the trick. David's paper is mainly concerned to show how an external epistemology like reliabilism might connect them (since, he argues, it has similar difficulties). To say p is justified, for a reliabilist, is to say it was formed by a reliable process (that is, a process that produces mostly true beliefs). Reliabilism

> seems to avoid construing the epistemic goodness of an individual justified belief merely as a constitutive means to the truth-goal; and it seems to take seriously the idea that truth is the epistemic goal.[20]

David's twist on the justification-truth connection is that the *process* generally produces true representations, and so believing the results of the process furthers the truth-goal. But that seems to be the same as saying that believing each individual result of the process furthers the truth-goal, even if a particular result of the process is false.[21]

It is a trick familiar from rule utilitarianism. The rule utilitarian does not like the counter-intuitive results of act utilitarianism, such as the occasional permissibility of murdering innocents. So, roughly speaking, she points to rules (like processes) that reliably bring about happy results—rules like "don't murder". You can follow this rule with happiness as your goal, even if individual instances of following the rule bring about less happiness overall. Following the rule is a means to good consequences, but following the

[20]David (2001) p165.

[21]This is still not enough, David argues; the result is subject to diachronic worries again, in determining the reliability of a process. David suggests replacing actual past and future results with subjunctive results—the process *would* now reliably produce true beliefs. This too seems to have problems, though; in what sense would they? —Even in a demon world?

rule is not the same thing as maximizing the good. The right and the good have been fallibly connected just as justification and truth are for the reliabilist.

Of course, if this analogy is right, reliabilism inherits the problems as well as the advantages of rule utilitarianism. In particular it is hard to see why following the rule is a good thing even when what justified the rule in the first place fails to hold. Why is it right not to murder the innocent person even when it is not good-maximizing, given that the right is determined by the good? Similarly why is it epistemically good to believe the false result of a reliable cognitive process?

At any rate suppose answers to these questions can be given. Though David fails to mention it, the same kind of route is available to *internal* epistemologies. Suppose we devise epistemic procedures that (as a rule) lead to true beliefs. Perhaps these procedures could be instantiated as cognitive mechanisms. And suppose our success at following these procedures is cognitively accessible. Then we have an epistemic aim that is importantly connected to truth, and that does not fall prey (when properly tweaked) to the problems with diachronic or constitutive accounts. But to count as an *internal* alethic epistemology, it is not enough that these procedures reliably bring about true beliefs (perhaps it is not even necessary). Just that part of the plan is familiar reliabilism. To have truth as a final internal aim is to be able to adjust ourselves toward the goal of truth, and to have our success in doing so cognitively accessible. This is a tall order—in fact, seemingly impossible given skepticism.

4.1.6 Lormand's responsible searching

But success may come in different degrees, and these degrees of success may have different dimensions. Some of these dimensions may be cognitively accessible while others are not. That is a crucial insight of Lormand's paper on *responsible searching* for truth.[22]

[22]Lormand (2001).

Lormand tells the story of Renée O'Cards, a reformed Tarot reader who is tired of believing whatever it is pragmatically effective to believe. She now has only one aim: to get all and only true beliefs. This aim is not all-or-nothing, of course; it can be more or less satisfied. And she does not just aim to *have* truth—in particular, "happy-go-lucky" truth would not count. She wants to *find* the truth through her own doing, via sub-aims that if satisfied would satisfy the main aim. In this way she will be "epistemically responsible". And, unlike the accounts considered above, "Renée doesn't take epistemic responsibility as a *means* to Truth." But

> her aim for epistemic responsibility is still "based" on her aim for Truth in a second sense: her aim for epistemic responsibility depends on the *existence* of her aim for Truth. She wouldn't have an aim for rational responsibility unless she had the aim for Truth.[23]

And to find something, you have to search for it. So to the extent that Renée is able to tell good truth-searches from bad truth-searches, she can have an internal aim that is based on truth.

Renée recognizes that the connection between her internal search and the goal of truth is a fallible one. She cannot distinguish good truth-searches from bad truth-searches based on which ones get her truth, since (as her skeptical intuitions demand) she has no independent grip on which beliefs are true and which false. But she *can* tell, perhaps, a good search from a bad search. A good search for any thing x will *cover* lots of candidates for x, and will *discriminate* the candidates into hypothesized x's and non-x's, and will eventually *stabilize* on some of the candidates as x's. (If a search does not even potentially stabilize on some candidates, nothing could ever be *found*, and so the proposed search is not a good one.) As Lormand puts it, when faced with skepticism,

[23]This quotation, and my paraphrasing previous, are from Lormand (2001) section II.

> One potential workaround for Renée—and the only one that occurs to me—is to pare away from the Truth-Specs issues about whether the relevant beliefs are true, leaving pared-down Specs about beliefs *independent* of their truth or falsity. For example, unless there are *some* candidate beliefs (whether true or false) that a method stabilizes on, it doesn't stabilize on *true* beliefs. ... And when beliefs that are covered, discriminated, or stabilized upon happen to be true—even if this is unbeknownst to Renée—satisfying these Belief-Specs satisfies the Truth-Specs.[24]

Let's call beliefs that meet the Belief-Specs the *covered discriminated stable* beliefs, or the *CDS* beliefs for short. So Renée can have internal feedback toward truth, while acknowledging skeptical possibilities, by having feedback toward good belief searches. That at least is a necessary component of finding truth. Lormand then argues that the best way to find truths is to use principles like simplicity (without which the search would not stabilize), analogy, parsimony, and a weak form of conservativism (that, as a tiebreaker, prefers old beliefs). Thus we have purely alethic defenses of some traditional theoretical virtues. This might help ground the notion of *evidence* for the evidentialist; Lormand's responsible search for a truth method could perhaps subsume evidentialism.

4.1.7 Dogmatism and truth

Lormand's is the best shot I know of for the alethicist. But there are reasons to doubt that it succeeds as an internal, alethic, *and* intuitive epistemology. In particular the bare principles of responsible searching do not seem to be enough to derive the theoretical virtues. From what I can tell, in fact, the *best* way to meet Renée's criteria for a responsible truth search is through an ultimate form of *dogmatism*, not through the traditional theoreti-

[24] Lormand (2001) section V.

cal virtues. The moment Renée works out her views on epistemic responsibility, she should simply believe fully all the candidate beliefs she presently has, reject all others, and never change her belief state again. To use the terms of Lormand's paper, it seems as though her metasearch (for effective belief-finding methods) should rate $\text{Conservatism}_{\text{virtue}}$ extremely highly, outranking all other scoring criteria. Dogmatism can work as a test for any belief ("do I already have this belief?"), so it has ideal coverage. It discriminates ideally, sorting only those beliefs she has into one category and all others into the other, with no messy intermediate degrees. And it is the most stable belief system imaginable. Renée would certainly *find* beliefs wonderfully with this approach, and the supposition is that finding beliefs is what Renée can best do to find true beliefs.

As it happens, Renée is not allowed this method, because

> Renée has a core ground rule for prideful epistemic responsibility. It constrains her cognitive starting point. She aims "carefully to avoid precipitancy and prejudice" where she "cannot exclude all ground of doubt." ... I will suppose that [except for beliefs of the form "I think p"], Renée is *unwilling* initially to accept contingent candidate beliefs as (probably or actually) *true* or (probably or actually) *false*. She will not accept such candidate beliefs without epistemically responsible reasons, and at the start she doesn't have such reasons.[25]

But wait—at the start, Renée *does* have an epistemically responsible reason to be a prejudiced believer: because doing so gets her covered, discriminated, and stable beliefs! (We are supposing that for a belief to be CDS is an epistemically responsible, alethic reason to accept it.)

So what is the motivation for Renée's core ground rule against prejudice? She seems

[25] Lormand (2001) section II.

to be just shooting herself in the foot with it. Lormand's view should explain why Renée cannot avail herself of an obvious and ideal method for finding beliefs. Without such an explanation, we are left wondering what work the prejudice rule does. Why is Renée more likely to find truth if she abandons her current potentially CDS beliefs and sets off on a protracted quest for *other* potentially CDS beliefs, given the supposition that all she can do to aim at truth is to aim at CDS beliefs? Of course *intuitively* dogmatism is a worse route to truth. But an internal alethic epistemology should either explain this intuition, or perhaps explain it away (and endorse dogmatism).

C. S. Peirce, the classic pragmatist, can explain our intuition against dogmatism—but naturally it is not an explanation that will be any help to Renée, since it is not alethic in nature. It is a curious coincidence that just like Renée, Peirce is interested chiefly in stable ("fixed") beliefs as an internal aim, having abandoned the direct aim of true beliefs.[26] And Peirce considers dogmatism the first, most obvious way to achieve it. In fact he reminds us that "this simple and direct method is really pursued by many men."[27] But he claims that dogmatism often will not work in practice for these men, because the opinions of others "will shake his confidence in his belief." A typical man will have pragmatic reason to reconsider:

> This conception, that another man's thought or sentiment may be equivalent to one's own, is a distinctly new step, and a highly important one. It arises from an impulse too strong in man to be suppressed, without danger of destroying the human species.[28]

And once pragmatic success is at issue, dogmatism can generally be dismissed more sim-

[26] See Peirce (1877), especially the second paragraph of section iv. We will look more at this coincidence in section 4.2.1.

[27] Peirce (1877) p68.

[28] Peirce (1877) p70.

ply than that. Even without the influences of society, a dogmatist will not find life easy. If he

> should resolutely continue to believe that fire would not burn him, or that he would be eternally damned if he received his *ingesta* otherwise than through a stomach pump,[29]

he simply will not be pragmatically successful, to put things mildly. And if these "inconveniences" somehow pale compared to the peace of mind that comes from stubbornly sticking to one's beliefs, then Peirce throws up his hands and says "I do not see what can be said against his doing so."[30] Peirce can give only pragmatic reasons to shun dogmatism. But Renée has managed to purge herself of such practical impulses, and so the Peircian route is not available to her.

Perhaps lack of prejudice *follows* somehow from Renée's insistence on taking epistemic responsibility for her beliefs. For example, when Renée is at the stage of considering methods for filtering beliefs for coverage, she reasons that

> any unevaluated filtering process is prejudiced, and if it nevertheless succeeds in filtering out false candidates, this is a matter of epistemic luck which Renée proudly aims to avoid.[31]

Here is a suggestion, then, about how prejudice and epistemic responsibility are tied together: if the prejudiced beliefs happen to get things right, it is just a matter of luck, and not hard-earned epistemic reward.

But this is puzzling. Suppose Renée starts off on her interminable search instead, first doing the meta-search (but rejecting the obvious virtues of conservatism on anti-prejudicial

[29] Peirce (1877) p69.

[30] Peirce (1877) p69. I admit that this is a somewhat tendentious reading of Peirce, to suggest that only pragmatic reasons move one from the other methods of belief.

[31] Lormand (2001) section VI.

grounds), and then carefully evaluating each and every candidate belief for simplicity, parsimony, and the like. And suppose she starts to amass a pile of CDS beliefs. Suppose further that this pile of beliefs turns out to be true. How is she any *less* lucky than if she had just *started* with true beliefs? I'll grant that she is responsible for having gotten CDS beliefs. But she would have been just as responsible for having gotten CDS beliefs through the dogmatic method. Meanwhile their truth seems no less lucky. As Renée is aware, she might be deceived by demons. Renée has already decided that given these problems, all she can do is aim for CDS beliefs, and hope for the best. That is the most responsibility she can take.

And in fact in terms of avoiding luck, too, dogmatism seems to win out. One concern of Renée's in avoiding luck is that she should not rely on the search taking too much time, because she might unluckily halt by, say, dying—this is why she decides to tolerate nonexhaustive generation.[32] The dogmatic method is practically instantaneous, while the method Renée mysteriously chose takes an indefinite and possibly infinite amount of time. Any number of search-threatening ills might befall her while she is undergoing that whole procedure. If any CDS beliefs will do—and again, why prefer later ones over the earlier?—better to get them quicker. Otherwise you risk not *finding* beliefs at all, let alone true ones.

From the assumption that an aim for CDS beliefs is an aim for truth, there seems to be no way to avoid the conclusion that dogmatism is the best way to aim for truth. This puts the responsible search method in an unintuitive position. We are intuitively against dogmatism, I think, for the same reason we are skeptics: we are reluctant to give up revisability of our beliefs. But it turns out difficult to explain this reluctance as the direct result of an interest in *truth*. The fear that our initial beliefs might be false is not enough to explain resistance to dogmatism, because any moderate realist will feel the same about any

[32]Lormand (2001) section VI.

subsequent beliefs. Why then should we move from one set to the other—what *internal* reason does a thinker have? Peirce has an easy answer to this question, but the alethic epistemologist does not.

4.2 Should we aim at truth?

Suppose, though, that all the foregoing considerations are wrong, and we *can* aim for truth in a meaningful way. Still, it is not so clear to me that we *should*. This may seem like an incredible claim, but I am not the first to make it; it is part and parcel of the epistemic pragmatist's outlook. Epistemologists who work with an eye toward artificial intelligence, it is worth noting again, tend this way.[33] But there are also more general arguments from epistemic pragmatists. Stephen Stich was even bold enough to give a chapter of his book the title "Do we really care whether our beliefs are true?"[34]

Like the title of section 4.1, the title of this section, too, is misleading. Unlike Stich, I think we *should* aim for truth—but only for its instrumental value. On my construal of epistemic pragmatism, true beliefs are like money. First, like money, true beliefs are excellent tools for obtaining what we desire, though they are neither sufficient nor necessary in this regard. Second, like money, they have no intrinsic value of their own. And third, I claim the instrumental value of true beliefs is so nearly universal that it is often *mistaken* for intrinsic value—just as Ebeneezer Scrooge mistakenly thinks accumulating money is its own end.

My first argument for these claims is a closer look at the odd result that both Lormand's Renée and a pragmatist like Peirce come to essentially the same aim for stable beliefs. If both form beliefs in the same way, then how does this difference *matter*? This question is closely related, I claim, to the typical pragmatic reaction to the skeptical hypothesis.

[33] See footnote 29 on p29.
[34] In Stich (1990).

Next I consider arguments specific to BDC against the intrinsic value of true beliefs, and then more general arguments to the same effect. Finally, since he is a leading exponent of epistemic pragmatism (though of the external kind), I spend some time discussing Stich's views on the intrinsic and instrumental value of truth. Though I disagree with much of his argumentation, I share his conclusion that true beliefs do not have intrinsic value. On the other hand I take issue with his claim that true beliefs also lack instrumental value—though I do share with him one concern in this area.

4.2.1 Skepticism and differences that matter

In section 4.1.7 we saw that Peirce the epistemic pragmatist and Renée the epistemic alethicist end up advising the same epistemic behavior based on two quite different motivations. A natural pragmatist would be inclined to ask, "so then what's the *difference* between them?" Or, more carefully, "why should we *care* about true beliefs, over and above caring about settled opinion?" This is a species of the standard pragmatist question "why does [some difference] *matter*?"—a question that I (as a pragmatist) think well worth asking.

For example, whenever I teach skepticism to undergraduates, at least one student in each class will eventually ask a closely related question. "Okay," the student will say. "Maybe I *am* [dreaming / deceived by a demon / a brain in a vat / in the Matrix / whatever]. So? What difference does it make?" Sometimes it's a bright student, and sometimes it's a surly, practical one tired of these "philosophical" problems. And to be honest, I never know quite how to answer those students. In fact I was the one raising the same question as a student in my own introductory class, and to my mind I have never heard a satisfactory answer. (That is not to say that I think skepticism is uninteresting or not worthwhile as a problem, though, as we will see.)

I do try my best to engage the students in the issue. One approach is to suggest that it *would* make a practical difference—a difference in action—if I thought some skeptical hypothesis were true. If I thought I were in the Matrix, say, I would quit my studies and go try to find Larry Fishburne, so I could get out. But astute students will notice my trick, and point out that the skeptical hypothesis gives me no reason to think I am in the Matrix—it just undermines reasons to think I am *not*. And without a reason to think I *am* in the Matrix, I should not go find Larry Fishburne.

I might then try to turn this argument around: according to the skeptic, I similarly do not have positive reason to believe I am in a "real world". And without such reason, maybe I should not bother to eat any more. But the reason for action in this case is different. If I just stop eating, on the assumption that I have no positive reason to do so, I will (I suppose) start to feel hungry, and I do not like that feeling. Sure, maybe I am not *really* hungry, because I am in the Matrix—but it will *seem* that way. The appearances give me *prima facie* reason to eat in a way they do not give me *prima facie* reason to go find Larry Fishburne. Whether I am a Berkeleyan idealist, or convinced a demon is deceiving me, or a regular old realist, I am going to try to get lunch after class.

Another thing I will do, no matter which of these metaphysical theses has a hold on me, is try to make sense of the perceptions and impulses that percolate into my mind, and act to adjust them in an apparently satisfying way. I am likely to do this in part by fitting particulars into generalizations, and by perceiving regularities, and by inducting from them. I could do all this and still be genuinely agnostic about what (if anything) underlies these apparent regularities. What the hell, I'd figure; all I can do is act according to appearances. It seems to me that when C happens E follows; it seems I can bring about C; it seems I want E—so I will bring about C. These things I will do whether I am an idealist or a realist.

So skeptical worries are of the type that seem to make no *difference*. This is a familiar pragmatist point; Peirce accuses Descartes of engaging in doubt that is not "real and living",[35] and the insouciant undergrad is accusing me similarly. Of course the undergrad will have several "real and living" doubts—about whether he will get a good grade in my class, or whether his girlfriend is angry at him, or whatever. And these he will try to settle. He will seek justification, broadly construed, for many of his beliefs. This attempt to settle doubt looks most naturally to a philosopher like he is seeking the *truth* on these matters, and yet his nonchalance about skepticism suggests that the truth of his situation is not of such great interest to him after all. (He is apparently not so bothered, in at least some sense, that his girlfriend may just be a figment of the Matrix.) A pragmatist will defend the consistency of his view, though. To a pragmatist, what looks like a search for truth just is a search for settled opinion.

And when alethic philosophers suggest that his search should be for something *more* than mere settled opinion—namely truth—it is not clear what recommendation they are making. How, the undergrad might reasonably ask, would the search for truth look different? In this way we are back to looking for a difference that matters between Renée's method and Peirce's. Richard Rorty opens a paper on the matter this way:

> Pragmatists think that if something makes no difference to practice, it should make no difference to philosophy. This conviction makes them suspicious of the philosophers' emphasis on the difference between justification and truth. For that difference makes no difference to my decisions about what to do. If I have concrete, specific, doubts about whether one of my beliefs is true, I can only resolve those doubts by asking if it is adequately justified—by finding and assessing additional reasons pro and con. I cannot bypass justification

[35] Peirce (1877) p68.

> and confine my attention to truth. Assessment of truth and assessment of justification are, when the question is about what I should believe now (rather than about why I, or someone else, acted as we did), the same activity.[36]

Peirce, of course, had similar instincts. Compare this quotation:

> ... the sole object of inquiry is the settlement of opinion. We may fancy that this is not enough for us, and that we seek not merely an opinion, but a true opinion. But put this fancy to the test and it proves groundless; for as soon as a firm belief is reached we are entirely satisfied, whether the belief be false or true. And it is clear that nothing out of the sphere of our knowledge can be our object, for nothing which does not affect the mind can be a motive for a mental effort. The most that can be maintained is that we seek for a belief that we shall *think* to be true. But we think each one of our beliefs to be true, and, indeed, it is mere tautology to say so.[37]

But what exactly does it mean to say that there is no *real* difference between aiming for truth and aiming for justification?

As Rorty suggests, the starting point for the pragmatist is that a difference has to have some import before it merits attention. Rorty and Peirce suggest that for the proper import, it must be a difference in "practice", or in a habit of action. Strictly speaking, though, the difference between committed idealists and realists does show up in their actions—for example, idealists will go about propounding idealism, and realists won't.

Presumably what the pragmatist really wants is a difference in action that *matters* somehow. The anti-skeptical undergraduate is a budding pragmatist because he is in effect asking why he should *care* about the problem of skepticism. The right way to approach

[36] Rorty (1995) p281.
[37] Peirce (1877) p67.

an answer, then, is to show how his life could potentially be better if we had an answer to the question. Perhaps the difference the pragmatists demand, then, is one in *instrumental* success. Plausibly doubts only have the telling Peircian itch when the uncertainty is potentially harmful to us, and we sense we would be *better off* with a settled opinion. (My "doubt" about whether I have hiccuped an odd number of times in my lifetime does not give me that itch, for example.) My pragmatic undergrads do not feel this itch when it comes to skepticism—and, most of the time, neither do I.

Suppose, then, that Peirce's and Renée's epistemologies end up making the same recommendations for cognitive behavior. If so, then is there really a difference between them? Or more carefully, is the difference between them worth *caring* about? Rorty's paper addresses this issue in a response to Crispin Wright. Wright argues that justification and truth are distinct norms for thinking, since "although aiming at one is, necessarily, aiming at the other, success in the one aim need not be success in the other."[38] The suggestion is that norms can be "prescriptively coincident", in Rorty's words, but not "extensionally coincident". I take this to mean: maybe two separate norms can tell you to *do* exactly the same thing, but have different satisfaction conditions.

Consider Adam, the ancient archer.[39] Adam took up archery to please Artemis, who he heard was into that sort of thing. But he knows that like all the gods, Artemis is fickle, and may no longer be interested in archery. In fact atheists and other heathens seem to think there may not even *be* an Artemis. But Adam's best guess is that there is an Artemis, and that she likes archery and good archers. And so aiming for the bullseye depends on his aim to please Artemis. He would not be an archer (suppose) were it not for his beliefs

[38]Rorty (1995) p288 quotes Wright's *Truth and Objectivity*. Strictly speaking, Wright was discussing a distinction between warranted assertability and truth (not justification and truth) as norms for "statement-making practices" (not thinking).

[39]This comparison is based heavily on a *Euthyphro*-like argument in Rorty's paper, for which Rorty is in turn "heavily indebted" to Bjorn Ramberg and Barry Smith.

that, at least, success at archery *might* please Artemis. Similarly, Renée would not be into settled beliefs if it were not for her belief that getting such beliefs at least might involve getting true beliefs.

Next to Adam in the archery competition is Anna. As Peirce is to Renée, so Anna is to Adam. Anna is also an ancient archer—an atheist one. Where Adam aims to please Artemis, Anna aims only for her arrows' accuracy. We can suppose that these norms are prescriptively coincident. Suppose further that Anna is right, and there is no Artemis. Thus their norms are not extensionally coincident. The satisfaction conditions for Anna's norm involve hitting the bullseye, while Adam's norm has no satisfaction conditions at all. He cannot do *anything* to please the non-existent Artemis.

You might ask: how is it that a norm with no "extension", to use Rorty's phrase, can have a "prescription"? How can such a norm tell us to do something? We could just as well say, if there is no Artemis, that there is no such norm as "please Artemis" at all, or that its prescriptive force is also empty. But the idea, I believe, is that we are discussing an "internal" norm—understood, for our purposes, as a conation of some creature with both causal effects on the creature's behavior (the prescription), and a representation of how things would be should the conation be fulfilled (the extension). Only understood this way can the "prescription" and "extension" of the norms reasonably come apart. Understood this way, though, we are left with the thought that Adam's and Anna's norms have the same effects on their behavior (both external and cognitive behavior). Both norms cause their possessors to practice archery, enter tournaments, aim carefully, and the like. So again, what's the big deal if one of them happens to have a different representation of the *success* conditions of their actions? Why might that be an important difference?

One possible answer on behalf of Adam and Renée: the norms come apart counterfactually. If Adam were to believe that Artemis is impressed by bowling instead, his norm

would lead to different behavior from Anna's. Similarly, I suppose, Renée might come to believe that having true beliefs does not involve having settled ones. She might miss Tarot, and come to believe that it is the best path to true beliefs. But the case is not as straightforward as Adam's, because the connection between settled beliefs and true beliefs is supposed to be analytic. Renée is given all the *a priori* truths (roughly speaking) up front, and it is from these that she derives her sub-aim for settled beliefs. Renée *could* of course search for truth in the Tarot, as many do—but your standard alethic epistemologist would think she is going about it wrong. And, it is important to notice, a standard pragmatic epistemologist such as Peirce would feel the same way (and further he might have better explanations as to *why*). Intuitions about good and bad ways to seek truth correlate with intuitions about good and bad ways to seek stable beliefs, while intuitions about how to be a good archer might not correlate with intuitions about pleasing Artemis.

Here is another possible answer on behalf of Adam and Renée: the stricter satisfaction conditions inspire more fervent dedication to the prescriptions of the aim, and that could make a difference. Adam's inspiration is a goddess, and perhaps that is more likely to rouse him out of bed for practice on a cold morning than Anna's merely mortal concerns. Similarly Renée will not settle for mere personal pragmatic comfort the way Peirce will. Rather she has noble, disinterested Truth to sustain her. Peirce may be content to be the victim of demonic deceit, but she aims for something higher. Of course Adam and Renée will not do anything differently in their pursuits, by our supposition that their aims are prescriptively equivalent. But maybe they will be more adamant, and maybe that is enough to make a difference that matters. Peirce might be relatively lazy because his stakes are lower, the supposition goes, and that risk of laziness is why it is important to pursue truth rather than simply stable beliefs. The inspired ones are more likely to *do* the things prescribed, or something.

But Peirce can be just about—maybe not *quite*—as harrowed by fear of deceit as Renée. Though full-blown skepticism does not scare him, he frets that he might be seriously deceived. That is, he worries that a great many of his beliefs may require serious revision. And these intuitions come from noticing that even beliefs that seemed secure have had to be drastically revised in the past. Genuinely stable beliefs, too, are very difficult to attain; Peirce's aim has satisfaction conditions that are *almost* as strict as Renée's. This reminder that no belief is safe keeps him as vigilant as the next person. He has, after all, *pragmatic* reason to remember that the search for stable beliefs is a demanding and difficult one. If he gets lazy, we suppose, the stability of his beliefs will falter, and this will cost him in pragmatic terms.

I take this to be the "cautionary" use of truth that Rorty endorses. To Rorty, one aspect of the notion of truth is pragmatically useful. As he puts it,

> the difference between justification and truth is one which makes no difference except for the reminder that justification to one audience is not justification to another . . . [40]

In other words, what looks justified today may not look that way tomorrow. So (even the pragmatist can say): keep on your epistemic toes.

Now I say the satisfaction conditions for Peirce are *almost* as strict, and he can be *almost* as proudly scornful of deception as Renée. The difference is in cases where nothing can be *done* about the deceit. Peirce and Renée could get ideally stable beliefs from experiences fed them by a demon, and Peirce's norm could be satisfied in this circumstance while Renée's could not. Of course Renée cannot *do* anything more than Peirce can about such cases. But she is welcome to worry, if she wishes, about things completely beyond her control. She apparently feels this is a difference that matters, but a pragmatist like

[40] Rorty (1995) p300.

Peirce or myself will have a hard time seeing why.

4.2.2 BDC and truth's value

If truth is of intrinsic value, though, then perhaps there is something like a *moral* difference between seeking settled opinion and seeking true beliefs. As Kant's distinction would have it, perhaps seeking truths is to act *from* duty to what is of intrinsic epistemic value, while seeking settled beliefs would be to act merely *in accordance* with the duty to truth.

If, that is, truth has intrinsic value. But BDC implicitly has a picture of value according to which the aim for truth is only instrumental. For the "rational" desires, just like the rational beliefs, are those endorsed in a state of belief-desire coherence. This coherence in turn depends at bottom on the default desires that come hardwired with a creature for obtaining its functions. Furthermore what is of value, plausibly, is the fulfillment of these rational desires. (The actual rather than the apparent fulfillment, that is.) And yes, a creature could form a *rational* desire to obtain true beliefs, it seems. But according to BDC's picture of learning and internal aims, this desire would be learned in order to further a proper balance of the creature's other, more basic wants. In other words, truth is only a rational desire insofar as it does contribute to (enough of) these more basic wants. In still other words, truth is of only instrumental value. If in some environment self-deception were the best way to achieve the balance of the creature's basic wants, then self-deception would instead be a rational desire.

But why place the source of value at the fulfillment of the creature's "fixed aims"—such as, in humans, eating and safety and affection and the like? On the functional account of creatures, the fixed aims are designed in service of a yet higher "basic" aim. In humans, for example, the fixed aims of eating and such are designed for fulfilling the basic aim of

genetic reproduction. Maybe we should say that fulfillment of a creature's fixed aims is also of merely instrumental value, in service of its basic aim. For humans, then, genetic reproduction would be of ultimate ethical value. That is a dubious ethical view at best, but what *principled* reason does BDC have for avoiding it? Why is value to be placed at the level of a creature's *fixed* aims?

If BDC cannot give such a reason, then its picture of value seems fundamentally flawed, and so there is no reason to worry about its proclamations on truth's value. Well, I do not wish to pretend that the ethics of BDC are worked out in great detail, but I can appeal to some basic ethical intuitions at this point. We think it is cruel to starve a cat, for example, but not cruel to keep it from having too many kittens. That is because getting food seems like the *cat's* want, while genetic replication does not. I suggested in section 2.2.4 that a creature's wants can come apart from the function that explains that creature's existence. And it seems a basic intuition that the wants of creatures are of ethical value, and not fulfillment of the creatures' designs (whether the design was intentional or unintentional).

So suppose it is right that fixed creaturely aims form the foundation for value. Still, the argument above presupposes that truth could be only a derivative desire from such aims. But perhaps for some creatures—maybe even for humans—truth is one of those basic aims. Then wouldn't truth be of non-instrumental, and perhaps even intrinsic, value? The resulting epistemology could be a hybrid of alethic and pragmatic epistemology; or rather, it would be a pragmatic epistemology that is *also* alethic. Creatures should try to play off their cognitive functions so as to obtain the best mix of food, affection, *truth*, safety, and so on. What is the motivation for assuming truth would not be a fixed creaturely aim?

First, and again, true beliefs might be and indeed likely would be a fixed creaturely aim. The optic nerve pathways, for example, are designed to represent the environment

accurately, and in this sense the creature has a basic aim to represent the environment accurately. Most reasonably-designed creatures in normal environments cannot help but try to get true beliefs. (They try to get true beliefs because it helps them serve further aims.) The question at hand is: would a creature likely have a basic *conation* for true beliefs, the way a human does for food? This conation, remember, is a representation of the goal state, something toward which the creature can adjust its cognitive dispositions (we are supposing, *contra* section 4.1). The supposition is that the creature is designed with this conation. But, first, this seems unlikely for humans. The truth-aim requires awareness of our cognitions, and awareness of a way things might be independent of (most of) them, and further a desire to have them line up. These it seems require a great deal of mental sophistication that we simply do not have as infants. If humans ever have the explicit desire for truth at all, it is because they painstakingly *learn* it—perhaps in the company of philosophers.

And second, as for the plausibility of *any* creature having a basic, non-instrumental desire for truth: well, we have already seen conceptual difficulties with the design of such an aim (in section 4.1.4). These considerations are not conclusive, of course. But even if there could be a creature with a basic desire for truth, the resulting hybrid epistemology would still be at root a pragmatic one. Yes, the standard of good thinking for this putative creature would include truth in the list of goods the creature basically wants, but for this creature too, good thinking is determined by its fundamental desires.

4.2.3 The intrinsic value of truth

But I do not need to rely on BDC for arguments against the intrinsic value of true beliefs; there are more general arguments as well. It is not easy to argue about the intrinsic value of something; accounting for taste is at least *difficult*. But in this section I can

consider common intuitions for truth's intrinsic value, anyway, and show why I think these intuitions are misguided.

As a warmup, let's start with a sociological point in the area: we humans do not *care* about most truths. We do not care what 785×36 is, generally—only in very specific circumstances. We similarly do not normally inquire into the geographical origin of our chalk, or the total volume of hair excreted by Richard Nixon on a given day in 1972, or the color of the lightest scale on a particular fish off the coast of Denmark, or countless other trivialities. It is only when the answers become relevant to our interests that we undertake such inquiries.

Of course this sociological fact does not establish that we should not care about such truths. Maybe we are just neglecting something of intrinsic value; we have been known to do it before. And surely, you might object, we *could* care about truth—as a final goal, or just insofar as it is truth. Well, I suppose we *could*; the point here is that we do not often appreciate what this would really look like. For one thing, if our final goal were to get truths, then we should not care *which* we seek; the bizarre facts above would be just as good as any other. Or if anything, we should choose those truths more easily attained, like the infinite truths of arithmetic. But in fact even in fields where pursuit of truth seems untainted by worldly desires—such as higher math—there are reasons to treat some questions as more interesting or important than others. What makes them more interesting or important? It seems to me the reasons we pursue some questions and not others reveals some *other* goal we have beyond attaining more truths.

Still, we do seem to have natural intuitions about the intrinsic value of truth. For example, people rarely say "ignorance is bliss" without a touch of irony. That is not because ignorance is never bliss—many times people *are* happier not knowing the truth. Instead it is because we cannot help feeling sorry for the person who is happier because he

is ignorant. There is something pathetic about such a life—apparently because something of value is missing. In the "bliss" case it is not happiness that is missing, so plausibly what makes that life pathetic is instead a lack of true beliefs.

But these intuitions are not a great test case. It is easy to assign value to *someone else's* true beliefs—as easy as it is to urge someone else to order the spinach salad while we get the cheesecake. It may be another matter to prefer the truth to our own happiness. Still, such intuitions are not uncommon. Suppose your belief that p is pleasant and important to you—perhaps the belief that you are good at what you do, or that a significant other loves you, or some such. And now suppose someone of impeccable credibility—like an angel, or maybe Lawrence Fishburne's character in *The Matrix*—offers you the *truth* on the matter of whether p. Of course the truth may be as you believed all along, but then again it may not. At any rate learning whether in fact p is unlikely to garner you any *more* happiness than you had before. Still, a common instinct—certainly mine—would be to accept the authority's offer, and get the truth. This seems to show that I would risk trading happiness for true beliefs, and thus that I value them even if they do not lead to other things of value for me. This in turn gives at least some good evidence that true beliefs are of intrinsic value for me.

Not so fast, though. There may be reason to think this is not really an attribution of *intrinsic* value. For example, perhaps my risked sacrifice of happiness is a strategic one, for more long-term happiness later. I may be gambling that in the long run knowing the truth will increase my overall happiness score, even if that score suffers in the short term. Better to find out now that I am caught in a loveless marriage than after ten years of deceit, I might reason—and better in terms of my happiness, not in terms of my knowledge.

We could try to control for this factor. Perhaps the impeccable source assures me ahead of time that the *net* effect of the truth, over the long run, will at best have no effect on me,

and at worst will make me much more unhappy. Perhaps I am unusual in this regard, but here is where I get off the boat—it is not at all obvious to me that I would, or should, learn the truth in that circumstance. But I recognize that others have clear intuitions for taking the truth even in such cases. I am inclined to suspect that people with those intuitions are simply not picturing the case vividly enough, though this is a risky claim toward which I can only make a gesture. Suppose instead of the offer at hand, the angel simply gave you a choice between a happier and a less happy life—and incidentally offered you a free, randomly-selected truth in compensation should you take the unhappy choice (maybe the one about the fish's scale color). I hope this decision would be at least more difficult. It may not look like the same case, because the truth on offer is no longer guaranteed to be an *important* one. But it is just this interest in only the "important" truths that leads me to suspect their value is instrumental. I would have to hear what could make one truth more important than another in a way that does not suggest its importance comes from its *use* to us. Of course I agree that a truth about famine relief is more important than a truth about the number of hairs on my head, and I would trade more happiness for the former than for the latter. But I do not myself see how this could be because the former has more *intrinsic* value.

Well, maybe truth does not have enough intrinsic value to outweigh happiness. But it could still have some small ϵ value, enough anyway to serve as a tiebreaker between two circumstances that are otherwise identical. In his introductory classes Louis Loeb is fond of asking undergraduates to consider the following choice: two courses are being offered on, say, ancient Etruscan history. Suppose you were guaranteed that the truth or falsity of the beliefs you gain from this class will not affect you ever. (Perhaps it is well known that armageddon is approaching or something.) Two old professors teach the separate classes. Both are equally entertaining, interesting, and such. But one has lost all academic

integrity, and simply spouts falsehoods for the fun of it. The other is pretty darn accurate about Etruscan history. Given that both classes will be equally engaging, wouldn't you pick the one teaching truths?

Well, perhaps; but again perhaps that is because it is difficult to picture the case vividly. Compare the following offer: you are to choose between life A and life B. In both you will have exactly the same amount of happiness, but in life A you'll be stinking rich. Be honest, now: doesn't this extra fact incline you a *bit* toward life A? And don't you feel, on reflection, that it *shouldn't* have such tiebreaking force? We just have an initial instinct to suppose that a life with more wealth will be a more comfortable and happy one. (Perhaps you are just not the money-grubbing type, and do not share this instinct. Then suppose then in life A you have more friends, or tastier food, or something else of immediate appeal to you, while told two lives still provide the same amount of happiness.)

But perhaps even picturing the case vividly, someone will attribute intrinsic value to the true belief. I can then only try to argue that such an attribution would be a mistake. Arguments against intrinsic value are not easy to make, but there *are* many clear cases of value misattribution, and we can see if the case at hand is like those. One possible approach is to show that valuing true beliefs intrinsically is a case of rule overgeneration, like valuing money intrinsically. Money is generally agreed to be of only instrumental value, but of instrumental value so pervasive and immense that it can easily look intrinsic to folks like Scrooge. He would rather have money than the good things it could buy. This is a clear instance of value misattribution—he is overgenerating the rule that one should try to get money, just as a kid overgenerates the rule for forming past tense in saying "runned". Both of these rules (get money, add 'ed') have good justifications, and both can be pretty smart rules of thumb to hold, because most of the time they work. But both have exceptions. Perhaps similarly the rule "get truth" is a wise rule of thumb that occasionally

misfires.

This is hardly an original strategy for arguing against intrinsic value intuitions; utilitarians tend to rely on it.[41] Yes, say many act utilitarians, in bizarre circumstances where net increased utility really would result, one person should be killed to save five. This is abhorrent to our intuitions, sure. But that is because our intuitions are often internalized rules designed to approximate the utility-maximizing act in most cases. That is, utilitarians have a strategy for explaining away intuitions about intrinsic value, and the strategy appears to apply equally well (however well that might be) for the intrinsic value of truth. We are so used to truth's usefulness that it is really hard to overcome seeking it even when we "know" it will be bad for us—just as a first-time skydiver hesitates to jump off of a plane into thin air, even though she "knows" it is safe.

On the other hand, this utilitarian-style response relies on the fact that cases where it is better to avoid truth are exceedingly rare; otherwise there is no such explanation for the deeply-ingrained intuition to seek it. But cases of bliss in ignorance are not so unusual, perhaps. Cognitively challenged people are often portrayed as naively content, Forrest Gump style, because they are simply unaware of life's challenges and travails. *If* there is something to this cliché, then we have another argument for the intrinsic value of truth: we all value truth enough to trade it for happiness, as witnessed by the fact that we do not line up for lobotomies.

Here, though, another utilitarian strategy gives the natural response: an invoking of higher pleasures. (Of course it is hardly a coincidence that these responses to apparent intrinsic value attribution spring from utilitarianism.) "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied," says John Stuart Mill, and he meant "better" in terms of happiness.[42] Some things of value—

[41] See for example Harman's discussion: Harman (1977) p155.
[42] Mill (1863) p148.

like appreciating Shakespeare, or solving a mystery game—may require true beliefs that themselves remain of only instrumental value.

This response would need fleshing out, since it is no good if possession of truth itself makes for a higher pleasure. Socrates is grimacing in apparent displeasure over some vexing issue. In virtue of what is Socrates "happier" in this state than the nearby smiling fool? If it is *only* that Socrates has access to more truth, then it is hard to say how this differs from attributing intrinsic value to his true beliefs. If on the other hand there is some kind of joy in Socrates' activity—if, say, he would rather attempt a solution on his own than be handed the truth by our angel—then there are grounds for saying that this activity makes for a happier life, and it is an activity the fool will rarely if ever enjoy.

This last possibility is not so far-fetched. And though none of these arguments is decisive, they make room to doubt the traditional positive reasons to think truth is intrinsically valuable.

4.2.4 Truth conditions and truth's value

Further, Stich has an argument *against* the intrinsic value of truth. Again, it is not easy to argue against intrinsic value (as Stich well knows[43])—someone could claim that some trifle has intrinsic value, and it is hard to say why it couldn't. Therefore Stich does not attempt to argue that no one could possibly value beliefs true intrinsically; of course someone could. His strategy is to show that concern with true beliefs is arbitrary, conservative, and parochial when compared to other possible ways to categorize and evaluate our beliefs. According to top-contender theories for truth conditions, for example, 'Jonah drank water' is true if and only if *that guy*, about whom the whale legends may or may not be true, drank H_2O.[44] Suppose these semantic theories are correct and do give the

[43] See Stich (1990) p132.

[44] This example is adapted from Stich's examples in Stich (1990) p115 and Stich (1991a) p182; of course these are borrowed in turn from Saul Kripke (1972) and Hilary Putnam (1975).

proper truth conditions. Still, we could say this sentence is *true**, for example, if and only if whoever best fits the cluster of properties we associate with 'Jonah' drank stuff that has the superficial qualities of water, so it could be true* if some George (who did live in a whale) drank XYZ. Stich asks: why be concerned about truth conditions instead of truth* conditions? The truth conditions comport with our intuitions, sure.[45] But why is that a point in their favor? Stich says:

> These alternative interpretation functions are not the ones sanctioned by our intuitive judgments. They strike us as wrong or inappropriate. But there is no reason to think that we could not retrain our intuitions or bring up our children to have intuitions very different from ours. … And there is no reason, or at least no obvious reason, to think that people whose intuitions diverged from ours in these ways would be any worse off. It is in this sense that the causal / functional interpretation function is not only limited but also *idiosyncratic*.[46]

This argument goes some way toward suggesting why we should hesitate to attach intrinsic value to true beliefs (not true* ones).

Still, it is not convincing as it stands. The argument only works to the extent that the truth conditions must agree with intuitions that are, at base, arbitrary. The arbitrary nature of our intuitions is not so obvious, however—surely not as obvious as Stich takes it to be. Yes, in the end what semantic theory we endorse will have to depend on some primitive stance of ours, such as what arguments we just find persuasive, or what premises we are willing to accept as primitive. You can express this point by saying, if you like, that in the end we must call on *intuitions*. This does not mean that the relation between propositions and the conditions that make them true is arbitrary. If we retrained our intuitions we could

[45] Not uncontroversially; I am not alone in suspecting that our intuitions prior to the *arguments* given by Kripke and Putnam were indeterminate at best, as I will suggest later.

[46] Stich (1990) pp116-7.

only artificially retrain them to prefer something wrong. Perhaps we could also train our kids to value pollution or fascism, but that is not to say our current preferences for clean air and freedom are arbitrary.

There is something more to Stich's argument, though. First, though our semantic intuitions may not be arbitrary in the sense that they are at least close to correct, the further decision to *value* the beliefs true by the these lights may be arbitrary compared to valuing true* beliefs. We can point to reasons for preferring freedom to fascism. What similar reasons are there for preferring true beliefs to true* ones, other than random tugs of the gut? If there is no good reason to prefer the former, then it is hard to see why the beliefs we deem true have any intrinsic value.

Stich's master argument seems to go like this. First, any acceptable meaning theory must agree with intuitive judgments of content on at least a substantive majority of cases. As he puts it,

> ... any theory of interpretation that assigns to all of my beliefs truth conditions pertaining to events in the Crimean War, or to events in my own brain, would be immediately ruled out of court if what the theory is supposed to do is explicate the intuitive ability that underlies our ordinary judgments about content or truth conditions.[47]

Second, these intuitions are at base arbitrary, cultural artifacts.

[47] Stich (1990) p105. Stich does allow that more revisionary meaning theories might find their support less in intuition and more in wider philosophical or scientific projects; his response:

> I am inclined to think that if an interpretation function does not cleave reasonably close to commonsense practice, it is hard to see why what the function is characterizing deserves to be considered a *truth* condition. But I don't propose to push the point, since it runs the risk of leading to a dreary debate about who gets to use the word 'truth'. Instead, I will simply stipulate that my skepticism about the value of true beliefs is restricted to accounts that assign truth conditions largely compatible with commonsense intuition (Stich (1990) p106).

Myself, I fail to see a clear distinction between deciding on "intuition" in Stich's sense, and deciding according to some wide reflective equilibrium that includes science and philosophy.

I am not sure what to make of this argument; I don't see why I couldn't argue similarly against the intrinsic value of *anything* that way. Take friendship, for example; we need an account that would explain the intrinsic value friendship has over similar relationships like friendship* (maybe fair-weather friends count as friends*). Any such account we give must surely accord on the whole with our everyday intuitions and judgments about friendship; an account of friendship that holds only between cups of pudding will be "ruled out of court". And these intuitions, it seems, are at base arbitrary cultural artifacts; we could train our kids to value friendship* instead.[48] Or so the argument would go—but it is not clear to me that we *could* so retrain our kids, and not at all clear that we *should*, and not a bit clear what disanology there might be between the cases.

Well, again, talk of intrinsic value of any kind is tricky. Perhaps we should refocus the debate on whether our choice of truth conditions over truth* conditions is *arbitrary*. Here Stich is clear:

> Why do we have these particular intuitions rather than those that would sanction REFERENCE*, REFERENCE**, or one of the others? The short answer, of course, is that no one knows in any detail just how these intuitions arise. But it's a good bet that, like other complex systems of intuitions such as those concerning grammaticality or morality or politeness, the intuitions in question are themselves culturally transmitted and acquired by individuals from the surrounding society with little or no explicit instruction. ... Whatever the explanation, it is clear that our intuitions do not result from a systematic and critical assessment of the many alternative interpretation functions and the various virtues that each may have. One way or another, we have simply

[48] One could say it does not count as a friendship unless there is some value attached to the relationship. But this would just shift the debate to whether any acquaintances count as *friends*. Similarly, of course, for truth—if true beliefs have to be intrinsically valuable just to count as true beliefs, then the question is whether we have any.

inherited our intuitions; we have not made a reflective choice to have them.[49]

As stated, I disagree. We *do* know how these intuitions arose—they are chronicled in philosophy journals over the years. The philosophical community *did* choose, reflectively, to focus on truth conditions instead of truth* conditions. Stich seems to buy a common line according to which Kripke and Putnam simply showed us something about our intuitions when presenting their meaning theories. According to this line, the philosophical community went the Kripke-Putnam route because it just *felt* right. In some sense, of course that is true. But that does not mean the choice was unreflective.

First, it is not clear that the philosophical community *had* intuitions about whether, for example, 'water' might refer to this bizarre XYZ stuff on Twin Earth. Probably the question had never *occurred* to most people. One could insist that the intuition was implicit, in the form of a disposition to form this opinion when presented with the puzzle. But this response from dispositions should be stated carefully. Did we also have the *intuition* that space was non-Euclidean, because we were disposed to choose a non-Euclidean theory when presented with the right kind of question? There is a very clear sense in which our intuitions back then were that space is Euclidean. In this common sense of "intuition", it seems the philosophical intuitions of the time, if anything, leaned *away* from Kripke and Putnam's views, and toward descriptivist theories.[50] After all, if Kripke and Putnam were just preaching to the choir, then it is hard to explain why so many people took interest in their work.

Stich suggests elsewhere that conceptual analysis is just "domestic cognitive anthropology."[51] But neither Kripke nor Putnam were out doing surveys about how people used words, asking unbiased questions about Twin Earth and farmer Aristotles in a double-blind

[49] Stich (1990) p120.

[50] And I assume we have been talking about *philosophical* intuitions all along—unless the claim is that everyone has intuitions about rigid designators and XYZ?

[51] Stich (1991b) p206.

study. Instead, Kripke and Putnam presented *reasons* for the claim that 'Jonah' and 'water' are rigid designators. They did not simply mention the possibility that these refer rigidly; instead they wrote whole papers with arguments for their views, just as you might argue for freedom over fascism. Kripke's argument consisted, for example, partly in showing that the competing theory could not make good sense of counterfactual claims. Even if in fact Jonah lived in a whale, it ain't necessarily so (or *a priori* so either). That is an intuition, sure—but it is hard to see in what sense it is an arbitrary one. Putnam's argument involved the nature of scientific inquiry; surely there is an important sense in which people of the 17^{th} century meant the same thing we do by 'water'. Otherwise incommensurability would make nonsense of scientific progress. Again, it is an *intuition* that science investigates objective stuff about which our theories might change—but again, hardly an arbitrary one. Kripke and Putnam showed us that *if* we want to preserve the counterfactual talk so central to philosophical discourse, and *if* we want to avoid incommensurability in scientific endeavors, we had better hope that 'Jonah' and 'water' refer rigidly.

At any rate there is good reason to prefer true beliefs to true* beliefs; the choice was not arbitrary, but designed to save deeper intuitions about counterfactuals and science. If Stich's beef is with those, then he should have at them—it strikes me as a losing battle. But suppose he did manage to convince us that we should prefer truth* conditions, or truth** conditions, or something. First, I think he could only so convince us by appealing to even deeper "intuitions" that overthrow the appeal of the Kripke-Putnam story. And second, in normal contexts his argument would not show that truth conditions are not as valuable as truth* conditions. Rather, we would take his argument to show that regular old truth conditions are simply different from what we thought—just as Kripke and Putnam showed that truth conditions are different from what the descriptivists thought, not that some truth** conditions are more interesting.

But this last point, I think, is where Stich is on to something. We do take agreement with our intuitions as a kind of evidence for what truth conditions are. We have an intuition that Aristotle might have been a farmer, so we conclude 'Aristotle' must refer rigidly, rather than being short for a description like "the greatest philosopher of antiquity". But maybe the naive descriptivists are right about 'Aristotle', and our intuitions are wrong—it is not true that Aristotle could have been a farmer.[52] This would be a highly counter-intuitive conclusion, yes. But why is that a mark against it? Why not simply start retraining undergraduate intuitions on possibility? Or, for that matter, why not simply leave our theory of truth conditions in tension with our intuitions?

If we had good reason to think the intuitions we have track our *interests*, that would be good reason to salvage them. That is, maybe if we did manage to retrain our intuitions, things would go worse for us—and that would be good reason to think that our intuitions are not arbitrary. This response gives up on the intrinsic value of truth, of course, and tries to justify our preference for truth conditions over truth* conditions on *instrumental* grounds. This line of argument is the topic of the next section. Meanwhile here is where I agree with Stich about the intrinsic value of true beliefs: it completely escapes me how there could be any other, non-instrumental reason to endorse our deepest intuitions about what makes a belief true.

4.2.5 The instrumental value of truth

Stich argues that true beliefs are those simply sanctified by a conservative, parochial, idiosyncratic, and ultimately *arbitrary* set of intuitions. Given the background of the previous section on the intuitional accounts of truth conditions, this section must address two avenues for response. First: is there reason to think that our semantic intuitions are a guide

[52] Of course true descriptivists, even back then, would have their own ways to accommodate this intuition, using clusters or the like. Modern descriptivist takes, incorporating rigidifying indexicals, are even better. I use the Aristotle case here, as in Kripke, just as a toy example.

to pragmatic success, and thus are non-arbitrary? Second: if so, would such non-arbitrary intuitions establish *true beliefs* to be of instrumental value? The second question may seem obtuse, but there is reason to think it has bite. As Stich puts it:

> Moreover, even if it could be shown that using the intuitively sanctioned interpretation function is especially conducive to survival or success, this would still not be enough to show that *having true beliefs* is more instrumentally valuable than having TRUE*...* ones. To show this, it would presumably have to be shown that the reason the intuitively sanctioned interpretation function is conducive to success is that it fosters believing the truth.[53]

Here is how I understand Stich's worry: it could be that the intuitions that lead us to prefer some truth-conditional accounts over others lead to success, but for all that are misleading about the correct truth conditions. Again, as I would put it, maybe truth conditions are such that Aristotle could *not* have been a farmer, even if this is horribly unintuitive—and even under the supposition that it is unintuitive because inconvenient for our further success.

But first things first. There is no need to wring our hands about this issue if we cannot show in the first place that our intuitions about truth conditions are likely to be of instrumental value, and thus non-arbitrary. I think we have already seen a bit of a positive argument for this claim, though: the semantic intuitions we got from Kripke and Putnam (for example) make sense of counterfactual beliefs and beliefs about the nature of scientific investigation. We are reluctant to give these beliefs up—and reluctant, I think, for pragmatic reasons.

It may still seem strange to speak about instrumental reasons to pick some accounts of truth conditions over others, but it is not so strange to talk of instrumental reasons to use a word one way over another—and these two kinds of reasons are at least closely tied.

[53] Stich (1990) p122.

Consider a "real-life" case of the stakes in word use. Paul Henle, writing on ordinary language philosophy, uses the example of the platypus. Upon discovery of the creature, biologists could agree that platypuses are warmblooded, egg-laying, lactating quadrupeds. It was once a further issue, it seems, whether to describe them as *mammals*. Presumably it was a fairly trivial issue, involving only convention-setting and such, but an issue nonetheless.[54] And though somewhat trivial, the decision to apply 'mammal' to platypuses was not completely arbitrary. Presumably, for example, it was out of the question to class them as reptiles—but it is not trivial to say why. I do not mean to say the decision was arbitrary; there were surely good reasons like "because it's not cold-blooded" or "because it has the wrong phylogeny". What is not trivial is to explain why *those* are good reasons. Probably, it is *important* to respect the deep distinctions between warm- and cold-blooded creatures. (That is: taxonomists wanted their classification system to respect such deep divisions.) There are likely to have been similar and more abstruse, technical reasons available to the biologists for not classing platypuses as birds, and so on.[55]

Granted, the stakes in these choices are not obviously earth-shattering in their significance; the cases where such subtle differences in word usage make any practical difference

[54]See Henle (1957), I think a very interesting paper. Henle's quotation on the platypus (from p221):

> Thus, if one were investigating someone's use of the term "mammal" before the discovery of the platypus, it would have been equally compatible with the rule "Apply 'mammal' to all haematothermal viviparous lactiferous quadrupeds" and "Apply 'mammal' to all haematothermal lactiferous quadrupeds." Even today one cannot tell from other people's use of the term whether the requirement of being lactiferous is necessary; and even if asked whether they would call a warm-blooded quadruped which didn't give milk a mammal, most people would not know.

[55]I can't resist two more quick examples of word meaning and instrumental use. First, modern difficulty over the word 'neutrino' in physics: according to the standard physics model, last I heard, neutrinos do not have mass. But the standard way to test for neutrinos has suggested that the few caught *do* have mass. Is the correct conclusion that there are no neutrinos (because the meaning of the term is closely tied to the standard model)? Or that neutrinos have mass after all (because the meaning of the term is closely tied to its detection processes)? Why? I do not suggest there is no answer to this question—my own inclination would be to say neutrinos have mass—I mean merely to point to the plausibility of the idea that such a decision is made on instrumental grounds. For another example, consider the *New York Times* article titled "New Discoveries Complicate the Meaning of 'Planet' " (Wilford (2001)), prompted by star-orbiting aberrations outside the normal vocabulary of astronomy.

are esoteric to be sure. But as in business, small bits of repeatable efficiency add up, and the more economy we can squeeze out of words the better. It is this principle that results in the streamlined devices of the important words, ones eminently adaptable to our purposes. This principle explains why there are magazine columns about proper word use. It is a common pragmatist view that words are tools for doing a job, and we would like them to do it as well as possible. This involves finding important concepts for the words to track, and this fine-tuning of concepts is a key job for the philosopher. It is hard to say why philosophy is useful without supposing conceptual innovations, examinations, and house-cleanings are useful—and it is hard to say why philosophy should continue to be done without supposing that it *is* useful. In summary, the presumption that it is important to use words well, tracking ever more useful concepts, is built into the business.

I have been suggesting that our intuitions about truth-conditions are roughly dependent on our intuitions about meaning, and that these are shaped by our intuitions about how we can use words to achieve our ends. The approach has been to show a common coincidence in meaning intuitions and usefulness, which might lead one to suspect that we have intuitions in favor of a semantic position only after seeing the potential use of it. This in turn amounts to arguing that our truth-conditional intuitions are formed by appealing to still deeper intuitions about what is good for us. But of course Stich could then question *those* still deeper intuitions. At some point the intuitional spade will be turned, and Stich can question the instrumental or intrinsic value of these bottom-level beliefs.

A common argument around the worth of these fundamental intuitions centers on natural selection—our base intuitions may have been selected, either genetically or culturally, for the achievement of our ends. If so, then in effect we can trust that our bottom-level intuitions are instrumentally efficacious, because if they weren't, we wouldn't be here. Stich responds to this argument with a barrage from the standard arsenal against the Panglossian

adaptationist. (In the end, I believe the evolutionary debate between adaptationists and their opponents is a tired and pointless one, best left to those with the antecedent political motivations strong enough to find it interesting.[56] Both sides agree that natural selection is a strong force driving change in genetic frequency, and both sides agree that natural selection is not the *only* such force. From this point, it seems, whether you call the evolutionary glass half-full or half-empty depends on background commitments and hobby horses.)

Here is Stich's argument, in a nutshell, against the (biological) evolutionary reason to think our intuitions worthwhile:

> There are many factors in addition to natural selection that drive biological evolution, including genetic drift, pleiotropy, meiotic drive, and others, and each of these is capable of leading evolution away from an optimal phenotype. Moreover, even when natural selection is the only force at work, it cannot be counted upon to select the best option among those available. Nor can it be taken for granted that the best option *available* is the best *possible* option—or even that it is a particularly *good* option.[57]

I grant all of this, except the very last phrase about selection's option being a *good* one. None of the rest argues against the adaptive success of our cognitive systems, and Stich has failed to establish the last part.

To see the point, imagine that Stich is not arguing here against the adaptiveness of our thought processes, but instead against the adaptiveness of the human eye. (This should be easy to imagine, since nothing in his argument against the adaptationist is cognitive-specific.) Here is an adaptationist response to such an argument: yes, maybe the selec-

[56] It is probably not coincidence that adaptationists like E. O. Wilson favor a view of ethics according to which what is right is what evolution has shaped us to view as right, while the vocal anti-adaptationist Stephen J. Gould was Marxist.

[57] Stich (1990) pp96–97.

tive forces that shaped the human eye did not find the *best* option out of those available. Surely it did not find the best option *possible*—otherwise we would not need telescopes and microscopes. But it seems extraordinarily implausible to suppose that the eye has been with us lo these millennia as a result of random genetic drift, pleiotropy, or other non-adaptive forces. And even if the eye were not selected for—even if the eye were a genetic epiphenomenon that served no function—we could *still* be confident that the eye, if not "particularly good", is at least *good enough*. It is good enough not to kill off the eye-bearing among us completely. (That is, it is at least good enough in the same way that the gene complex that sometimes leads to sickle-cell anemia is "good enough".)

Just the same point can be made in favor of the brain, of course.[58] It is extremely likely that the human brain was selected for because of the cognitive processes that it underwrites, which suggests that these cognitive processes are in a basic sense successful, just as the eye is. At any rate our extraordinarily complex cognitive processes are "good enough" not to have killed us all off, and the immense energy and other evolutionary costs with which our brains burden us suggest that they may even provide some decent, compensating service for us on occasion.

So I think Stich's argument against the adaptiveness of our cognitive system badly overgeneralizes. But even if our base cognitive processes are evolutionarily adaptive, this does not yet show that they are instrumentally efficacious—unless we happen to value gene copies intrinsically. Of course we might intrinsically value things like longer, healthier, and safer lives, and it is likely that such lives are well correlated with efficient gene copiers (in humans anyway). If so, we have *some* reason from evolutionary biology to think our intuitions are likely to bring about things of value to us. But when it comes to intrinsic values that run counter to genetic benefit—such as a desire to reduce population growth

[58] See Pinker (1997) for an eminently readable defense of the brain as a selected-for organ, though this should hardly need defending.

for the sake of the environment—the evolutionary argument is at a loss. If anything, base intuitions that result from selective forces would lead us *away* from attaining such non-adaptive values.

Obviously, though, our intuitions are shaped not just by evolutionary forces, but also by cultural ones. Stich considers this avenue, too:

> Of course, when the evolution of concepts is at issue, it is likely that the processes involved are more social than biological. We have only a very primitive understanding of the processes at work in social evolution, but what little we do know hardly supports the suggestion that cultural products always or typically evolve in an instrumentally efficacious direction. One would be hard put to make a serious case for generally increasing adaptiveness or utility in the evolution of clothing styles, manners, religious practices, syntactic structures, or political systems, to mention just a few.[59]

Again, granted: cultural products surely do not *always* "evolve in an instrumentally efficacious direction." And perhaps they do not even typically so evolve. As in the case of biological evolution, there are at least *some* cultural phenotypes, and maybe most, that are not selected for. But that does not mean that *no* cultural phenotypes are selected for. Just because eye color may not be adaptive does not mean eyes aren't. The cultural case is analogous: just because fashion may not evolve in an efficacious direction does not mean that no cultural products do. Stich has listed some of the least likely products to undergo productive cultural selection.[60] But consider the paragon of cultural products: science. Does Stich wish to argue that the cultural evolution of science does not track instrumental

[59] Stich (1990) p97.

[60] Though even these may have evolved toward some instrumental benefit, in my view; common clothing styles are certainly more utilitarian than Victorian fashions or ill-fitting animal skins, say; and political systems like democracy strike me as having shown remarkable progress since the common tyrannies of old.

value? This is enormously implausible. Even staunch Luddites would have to admit that science has given us at least a greater range of possible goals to achieve, even if some feel the new technologically achievable goals themselves—including besides nuclear armageddon, remember, also curing diseases, feeding the hungry, *etc.*—are not worthwhile. And science may not be the only such cultural product. Ethics is another cultural product, and plausibly it too has shown real progress over the years, helping make for a generally better world than the slave-driving, draw-and-quartering days of the past. It is even possible to make similar cases for the arts and, well, just about anything taught at an accredited university. (Instrumental efficacy would certainly help explain why people pay such high tuition.)

Some cultural products, at any rate, do evolve for the better over time. That leaves it an open question whether our semantic and conceptual intuitions are one such. Stich suggests that in such a claim

> one can almost hear … the ghost of J. L. Austin who held that there is much traditional wisdom and experience distilled into the categories and distinctions embedded in everyday language.[61]

Stich says this as though it were a *reductio* to have a view like Austin's, but I think Austin was right on this score. (Much as he gets maligned these days, Austin was no dummy.) There *is* wisdom in sorting the world into conceptual categories like [water] and [poisonous] and [tiger], and still more wisdom embedded in concepts like [number] and [atom]. Modern computers owe their existence to concepts like [effective algorithm]—hardly an "everyday language" concept, but no less evidence of conceptual progress. This unfolding conceptual evolution is often antecedent to the uncontroversial success of technology.

[61] Stich (1990) p96.

What is clearly wrong is to believe that ordinary language already embeds *all* concepts, or the *best* concepts. But even Austin never believed *that* crazy view. Here is the closest he comes to the Panglossian view often attributed him:

> ... our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the life-times of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of the survival of the fittest, and more subtle, at least in all ordinary and reasonably practical matters, than any that you or I are likely to think up in our arm-chairs of an afternoon—the most favoured alternative method.[62]

But even this passage does not say that our conceptual toolbox is ideal; it says merely that it has had many generations of refinement. We do not have "all the connexions men have found worth drawing" period; we have all those gathered over millennia of conceptual development. Later in the same paper Austin makes this explicit:

> if a distinction works well for practical purposes in ordinary life (no mean feat, for even ordinary life is full of hard cases), then there is sure to be something in it, it will not mark nothing: yet this is likely enough to be not the best way of arranging things if our interests are more extensive or intellectual than ordinary ... superstition and error and fantasy of all kinds do become incorporated in ordinary language and even sometimes stand up to the survival test (only, when they do, why should we not detect it?). Certainly, then, ordinary language is *not* the last word: in principle it can everywhere be supplemented and improved upon and superseded. Only remember, it *is* the *first* word.[63]

[62] Austin (1956) p182.
[63] Austin (1956) p185.

This strikes me as an eminently reasonable position, and the two passages together form an elegant statement of the claim that our linguistic and conceptual culture undergo a kind of cultural evolution with pragmatically effective results.

Against both the biological and cultural evolutionary arguments, Stich seems to mistake the claim that our intuitions are *a good starting place* for the claim that they are *ideal*. Only the former is required to defend their instrumental efficacy, and on both the biological and cultural front, Stich attacks only the latter.

Stich's other, non-evolutionary argument against the instrumental value of our meaning intuitions has the same main difficulty. "[I]n many cases," he says, "we already know that having true beliefs would not be the *best* way to achieve our more fundamental goals."[64] As an example he tells the story of Harry, who had a true belief about when his plane departed, so made his plane on time—and as a result died in a plane crash. Take an account of truth* conditions for Harry's mental states that is just like their truth conditions except 'my plane leaves at 8:45a' is true* if and only if Harry's plane leaves at 7:45a. Then Harry would have been better off with a true* belief than his true one, so true beliefs (beliefs whose intuitionally sanctified truth conditions are met) are not always the most instrumentally efficacious. Again: fine—but who said anything about "always", or "best"? We need only that true beliefs are *usually* good to have, as a rule. If we could somehow work it out that 'I want a gun' means *I want a lollipop* when bad guys say it, and *I want a gun* when good guys say it, that might make for more instrumentally efficacious truth* conditions. And if we could make screwdrivers that miraculously turned into bottle-cap openers when applied to beer bottles, that might make for a more instrumentally efficacious toolbox. But given our limitations in these regards, we do the best we can. Having '8:45a' mean, univocally, 8:45a is an effective tool for achieving time-coordination, even if meeting such a goal is

[64] Stich (1990) p122, emphasis his.

sometimes disastrous. In summary, meanings can only be set so that they manage to do the most help for the most people.

And though Stich is right that there is surely room for improvement, he hardly shows that our truth conditional intuitions are not instrumentally good. Perhaps Harry and his co-passengers would have been better off with true* beliefs (it is very unlikely, note, that many others would be, even if the deviation from truth conditions applies just to token representations of that day). But compare true** beliefs, which are not at all like true beliefs; in particular 'my plane leaves at 8:45a' is true** if and only if birdbaths lick antelopes. In the space of possible interpretation mappings that Stich wants us to explore, the vast majority will be utterly useless. Compared to intuitions that favor these random mappings, our intuitions favoring truth conditions are outstanding.

Stich considers a response something like this; he says

> It might be claimed that although there are special circumstances in which TRUTH**** is more conducive to survival, and other circumstances in which TRUTH***** is more conducive to love, nonetheless one would still be better off seeking plain old uncapitalized truth, since it did a better job *in general*, or *in the long run*. Well perhaps it does. But to show this requires an *argument*, and as far as I know, no one has any inkling of how that argument might go.[65]

On the one hand I have already suggested just such an argument. Basically it runs like this: our intuitions about truth conditions have done a fine job for us; our genetic and scientific success are in part a testament to them. Maybe our intuitions in this realm are not the best possible ones, but they are at least good enough.

Stich, though, could grant that our intuitions are okay, and still demand that we stop valuing them when there could be much better out there. But here, I think, is the real

[65] Stich (1990) pp123-124.

kicker: if we were convinced that another interpretation function would do a better instrumental job, we would not conclude that truth is not as useful as truth*; we would instead conclude that we have discovered something about *truth* conditions. Our deeper intuitions would lead us to revise what we thought were proper truth conditions. This, I suggest, is just what has already happened in cases like the Kripke-Putnam proposals, and for that matter any other past refinement of semantics.

So on the one hand I think Stich's argument against the instrumental efficacy of our intuitions is no good; on the other hand, though, I think there is something very suggestive in what Stich says. Remember we needed two steps to show that true beliefs are instrumentally valuable. First, we needed to show that our semantic intuitions track what is instrumentally valuable, and this I have tried to do. But second, we needed to show that therefore *true beliefs* are instrumentally valuable. And here, like Stich, I too have no inkling of how such an argument would go. Only *assuming* that our (useful) semantic intuitions track truth conditions do I have an argument against Stich to the effect that truth is of instrumental value.

4.2.6 Meta-pragmatism

We return, then, to an earlier problem: just because it is horribly counter-intuitive to suppose that Aristotle was necessarily a philosopher, why does it follow that the meaning of 'Aristotle' and so on are such that it is *true* Aristotle could have been a farmer? And supposing it is horribly counter-intuitive 'water' should include XYZ in its intension, why would that mean it is *true* there is no water on Twin Earth? In the previous section I tried to argue that there is reason to follow our intuitions, but nowhere did this rely on, or support, a claim that our intuitions are truth-tracking.

It could be argued that the instrumental efficacy of these intuitions is evidence of their

tendency to track truth. But I do not see how such an argument would go. Of course if one is a Jamesian about truth, and the truth just *is* the efficacious—well then, no problem. But few people, even among self-professed pragmatists like myself, are willing to go that route. I have already argued for why such hesitation is warranted: it is fundamental to a common conception of investigation of the world that the way the world is could, in principle, outrun our epistemic access to it, now and forever. Any belief we now have may need to be revised later. But without a Jamesian conceptual tie between truth and efficacy, I do not see how we can show any correlation between one and the other. Datum: my intuitions say Aristotle could have been a farmer, and so 'Aristotle' is a rigid designator rather than an abbreviation for a definite description.[66] To make for a correlation, we then need a datum of another kind: it is *true* that Aristotle could have been a farmer. How are we to get this, *besides* appealing to our intuitions on the matter?

In fact it seems we do take our intuitions as evidence for what the truth conditions of our representations are. And this is, upon reflection, a strange habit. Taking intuitions as evidence for *successful* interpretation functions is one thing, and for *truth conditions* another. I have only been able to argue for the legitimacy of the former practice. And notice that in these arguments I do not just claim that we *should* choose the more instrumentally efficacious semantics; I argue that in practice we *do* so choose, in following our intuitions. (I further endorse this practice.)

At times, in fact, this argument tempts me to a position of *meta-pragmatism*, patterned after Georges Rey's "meta-atheism". Rey argues that even those who profess to be theists usually are in fact, and despite themselves, atheists.[67] I wish to make a similar suggestion

[66] Actually *my* "intuition" is sympathetic to a view that proper names are a sort of description, really a function from contexts to extension, that includes a possible world indexical—thus explaining its rigidity.

[67] In summary, he tries to support his suspicion that "many people's belief in God is more like a pretense of belief, like what we bear to an entrancing story (or our parent's honesty or our spouse's fidelity) that we very much wish were true even though we know better." See Rey (2000).

about pragmatism—the vast majority of us, I suggest, are pragmatists at heart. We choose even our philosophical positions based on instrumental efficacy, and like some pragmatists we often confuse the effective with the truthful. Philosophical analyses are really attempts to hone the items in our conceptual toolbox for greater overall utility.

Many philosophers will find an urgent need to resist the possibility that they are, on some level, pragmatists. But why? Surely one cannot point to the disastrous *consequences* of being a pragmatist; such consequences would just be taken on board by the pragmatist herself as a bad sign for whatever version of the view has those consequences. If instead there are deontic-like reasons to avoid pragmatism, it is hard to see what they could be, given the difficulties for valuing true beliefs intrinsically.

4.3 Concessions to truth

I do not mean by these arguments, though, to suggest that talk of truth should be thrown out of philosophical discussions as useless blathering. Truth has its place among external evaluations of thinkers, for example, and in metaphysical discussions. There is even a reasonable *place* for truth in internal epistemologies, as I am about to explain. This chapter only argues that this place is not as central as has traditionally been held—that there is in fact plenty of room for a pragmatic approach to internal epistemology.

4.3.1 *A priori* beliefs and external aims

A few times now I have conceded that a creature may have an aim (or several aims) to gain true beliefs.[68] In fact, as I have said, it is quite likely. Creatures are hardwired to bring about *external* functions, and chances are good they will perform these functions best when they make accurate assessments of their environment. Thus they are likely to be hardwired to make such accurate assessments. BDC represents these basic computational

[68] In sections 3.2.2, 4.1.2, and 4.2.2.

dispositions as "default" thoughts; those thoughts have, in effect, *primitive* justification. Since they are believed and justified not due to learning from pragmatic feedback, but rather due to default constraints the external world imposes, we might say that these beliefs are internally justified based simply on their truth. Of course the story is not quite that simple; the justification of these thoughts is defeasible rather than foundational, according to BDC. So the internal justification is plausibly truth-based for the default thoughts only to the extent that pragmatic considerations do not override them.

Truths about a creature's own experiences, for example, are pretty well assured by the hardware. Any creature with perceptual inputs and sophisticated enough mental computation to process these inputs on higher levels will very probably have cognitive states justified in the "default" way simply due to the creature's wiring.[69] This does not imply that our beliefs about our experiences are incorrigible; we can misrepresent our own experiences as well as external objects, it seems to me. I just claim that experiential beliefs will, in undefeated cases, have a primitive justification that is probably a result of their truth. Again, this comports with the nearly universal dictum of internal epistemology, "trust, *ceteris paribus*, your senses."

Similarly there may be primitive "laws of thought" with default justification, such as the standard truths of logic. We may simply be wired to accept q when we accept p and $p \rightarrow q$. And the explanation for our being so wired may be in the truth-preserving property of such dispositions. There may be analogous explanations for our rejection of contradictions and acceptance of tautologies.

A creature's hardwired thought-forming mechanisms and resulting beliefs that it must at least initially take on face value we might reasonably call "*a priori*", or perhaps even

[69]Perhaps even a retinal cell impulse is a "cognitive state", and whether you want to call it a "true" one (when, say, the cell functions properly) is up to you; I would be more inclined to attribute truth only to cognitive states with content capable of misrepresenting, like beliefs.

"transcendental", for that creature. But neither phrase applies in their traditional senses, of course, since the beliefs forced upon the creature in this way may be false—as in the case of hallucinations. The creature may have internal reason to override the hallucination, and dismiss it; or the creature may simply be deceived. It may even be that the creature is *designed* to misrepresent its environment in systematic ways, for the good of its external goals. Humans *might* have such a failing when it comes to reckoning probabilities, for example.

If on the other hand some thought-forming mechanism is not basic to the creature's computation—if, that is, the creature can *learn* to compute otherwise—then it is hard to see how our acceptance of them relies on their truth. If we do not primitively shun contradictions, for example, then we must have learned to shun them. That means, I have proposed, that we adjusted our cognitive mechanism toward avoiding them in accordance with the feedback from some higher aim of ours. In other words, if we learned to shun contradictions, it is because they are ineffective for attaining our higher functions. The same goes for any thought-forming mechanism a creature can learn to form or unform. And importantly, it is only these *learnable* mechanisms that are on the table for internal, normative epistemology. We can of course censure a creature for possessing primitive, unchangeable mechanisms that are inept or not truth-conducive, but this would clearly not be a norm that fits any of the intuitions from internal epistemology.

Again, though, contradictory beliefs are probably ineffective for attaining our higher functions because their conjunction is always *false*. If the mechanism against contradictory beliefs is primitive, it is probably because it is such a great way to help represent the world—*any* world—truthfully that it came hardwired, without any loss of adaptability. If on the other hand it is rationally learned, it is because it appears to help fulfill aims. And it probably appears that way because it *does* help fulfill those aims, and it probably does help

fulfill those aims because it helps represent the world accurately. A creature that can learn makes internal adjustments according to internal standards, but all these internal standards are aimed at the external functions for which the creature was designed. Its notion of improvement, and thus the justification for its conations, derives ultimately from these external aims. Similarly, given that true representations are likely to be an aid to these aims, there is an important sense where justification for its cognitions derives ultimately from representing the environment truly. A creature badly deceived by a demon can still have justified beliefs according to the internal norm of BDC in part because that creature was designed to represent some similar environment *correctly*. In this sense the internal epistemic norm of BDC depends on an external one.

4.3.2 Pragmatic evidentialism

In addition to having hardwired mechanisms designed for truth, a creature can also have an explicit desire for true beliefs. And this can be a *rational* desire even on the pragmatic, BDC understanding of "rational". The sequence, to a pragmatist, might look something like this: I start with a lot of perceptual experiences and some basic learning mechanisms. I learn a bunch of useful regularities in my experiences that help me achieve my functions. If sophisticated enough I notice these regularities, and they give me pragmatic reason to suppose that there is something external to my mental states that underlies these regularities. (That is, this supposition spurs me to look for other regularities, which turns out useful.) I also notice that I often have to revise my beliefs in order to assure pragmatic success, and so I notice that I cannot just believe whatever I want—believing in some ways works better than other ways. In particular, if I am smart, I notice that beliefs formed from wishful thinking often are associated with failure at my pragmatic goals. This leads me to a notion of better and worse beliefs that does not have to do with what

I want to believe (though they still have to do with what serves my desires best). In this way I form the notion of an "epistemic" reason to believe rather than a "merely pragmatic" one. I class "because experiments demonstrate it" in the former camp and "because I'll get paid better if I believe it" in the latter. I notice that the regularities I perceive are both instrumentally valuable and largely independent of what regularities I hope to perceive, and so I consider the "epistemic" reasons to believe to derive from this mind-independent externality that underlies these regularities, which I call the *world*. Given pragmatic reason to suppose there is an external world, I now desire to believe in accordance with it. That is, I desire true beliefs.

It might even look, then, like my own pragmatic epistemology demands an alethic epistemology. Epistemic pragmatism might be "self-effacing" in just the way Derek Parfit worries that consequentialism could be.[70] But for one thing, as I have argued, it is not clear how the desire for truth differs (in practice anyway) from the pragmatic desire to have stable beliefs *à la* Peirce. To the extent they are different, anyway, it is clearer how the latter can guide me. And furthermore, just like Parfit's self-effacing consequentialism case, the ultimate justification for these explicit alethic desires, I propose, is pragmatic at heart. If so then the alethic desire is only justified insofar as it serves further pragmatic aims. In other circumstances—such as the world where desires come true as a law of nature—the "alethic" desires will not be so justified.

The metaphysics of BDC mirror this strange situation. According to my version of epistemic pragmatism, there is pragmatic reason to posit an external world that our mental states can get right or wrong. Does this mean BDC takes the side of the metaphysical realist? After all, it says there is a real world. Unlike the pragmatic view of James, truth is not linked conceptually to pragmatic success. And unlike Putnam's "internal realism",

[70]In Parfit (1984). It could be, Parfit supposes, that the consequence-maximizing thing to do is to get everyone to be deontologists.

no amount of (internal, pragmatic) BDC-style justification will ever guarantee getting that world right.[71] BDC's metaphysics looks like the traditional "explanationist" approach to scientific realism according to which the best explanation of our scientific success is that our theories are tracking an independent world. On the other hand, perhaps BDC is an irrealist view, since this supposition of an external world ultimately rests on pragmatic grounds. Or, put another way, the fact that an external world is the "best" explanation of our experiences is no indication of the *truth* of that claim. It is the best explanation because it appears the most useful.

Though metaphysics is not my specialty, I am inclined to agree with Rorty and Harman on this point:[72] the proper pragmatist response to the metaphysical question of "realist or antirealist?" is that this is a false dichotomy. The view is a pragmatic one, and leave it at that.

[71] See Putnam (1981).
[72] See Harman (1999) p93.

Bibliography

Alston, W. P. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives*, 2:257–299.

Alston, W. P. (1993). Epistemic desiderata. *Philosophy and Phenomenological Research*, 53(3):527–551.

Austin, J. L. (1956). A plea for excuses. In Urmson, J. O. and Warnock, G. J., editors, *Philosophical Papers*, pages 175–204. Oxford University Press, 1989 edition.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford University Press, 1984 edition.

Cherniak, C. (1986). *Minimal Rationality*. MIT Press.

Churchland, P. M. (1995). *The Engine of Reason, the Seat of the Soul*. MIT Press, 1996 edition.

Churchland, P. S. and Sejnowski, T. J. (1992). *The Computational Brain*. MIT Press, 1993 edition.

Clifford, W. K. (1879). The ethics of belief. In Feinberg, J. and Shafer-Landau, R., editors, *Reason and Responsibility*, pages 121–125. Wadsworth, eleventh edition.

Cohen, S. (1984). Justification and truth. *Philosophical Studies*, 46:279–295.

Conee, E. and Feldman, R. (2001). Internalism defended. In *Epistemology: Internalism and Externalism*, pages 231–260. Blackwell Publishers.

Craig, E., editor (1998). *Routledge Encyclopedia of Philosophy*. Routledge, online edition.

Crumley II, J. S., editor (2000). *Problems in Mind*. Mayfield Publishing Company.

Dalgleish, T. (2003). Information processing approaches to emotion. In Davidson et al. (2003), chapter 33, pages 661–673.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons.

David, M. (2001). Truth as the epistemic goal. In Steup (2001), pages 151–169.

Davidson, R. J., Scherer, K. R., and Goldsmith, H. H., editors (2003). *Handbook of Affective Sciences*. Oxford University Press.

Davies, M. and Humberstone, L. (1980). Two notions of necessity. *Philosophical Studies*, 38:1–30.

Dennett, D. C. (1981). True believers: The intentional strategy and why it works. In Crumley II (2000), pages 226–242.

Dennett, D. C. (1994). Language and intelligence. In Khalfa, J., editor, *What is Intelligence?*, pages 161–178. Cambridge University Press.

DePaul, M. R. (2001). Value monism in epistemology. In Steup (2001), pages 170–183.

Dretske, F. (1986). Misrepresentation. In Crumley II (2000), pages 329–340.

Ellsworth, P. C. and Scherer, K. R. (2003). Appraisal processes in emotion. In Davidson et al. (2003), chapter 29, pages 572–595.

Feldman, R. and Conee, E. (1985). Evidentialism. In Sosa, E. and Kim, J., editors, *Epistemology: An Anthology*, pages 170–181. Blackwell Publishers.

Fodor, J. A. (1987). *Psychosemantics*. MIT Press.

Gallistel, C. R. and Glymour, C. (1998). Learning. In Craig (1998).

Glymour, C. (1991). Freud's androids. In Neu, J., editor, *The Cambridge Companion to Freud*, pages 44–85. Cambridge University Press.

Godfrey-Smith, P. (1998). Semantics, teleological. In Craig (1998).

Goldman, A. I. (1986). *Epistemology and Cognition*. Harvard University Press.

Goldman, A. I. (1999). Internalism exposed. In Steup (2001), pages 115–133.

Haack, S. (1993). *Evidence and Inquiry: Toward Reconstruction in Epistemology*. Blackwell Publishers.

Harman, G. (1977). *The Nature of Morality*. Oxford University Press.

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

Harman, G. (1999). Pragmatism and reasons for belief. In *Reasoning, Meaning, and Mind*, pages 93–116. Oxford University Press.

Henle, P. (1957). Do we discover our uses of words? In Rorty, R. M., editor, *The Linguistic Turn*, pages 218–223. The University of Chicago Press, 1992 edition.

Hopkins, J. (1998). Freud, Sigmund. In Craig (1998).

Humberstone, I. L. (1992). Direction of fit. *Mind*, 101(401):59–83.

Jain, S., Osherson, D., Royer, J. S., and Sharma, A. (1999). *Systems That Learn: An Introduction to Learning Theory*. MIT Press, second edition.

James, W. (1896). The will to believe. In Thayer (1982), pages 186–208.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65:575–603.

Kelly, K. T. (1996). *The Logic of Reliable Inquiry*. Oxford University Press.

Kripke, S. (1972). *Naming and Necessity*. Harvard University Press, 1993 edition.

Lormand, E. (2001). How to (start a) search for truth. http://www-personal.umich.edu/∼lormand/phil/epist/how2search4truth.htm.

Lycan, W. (1988). *Judgement and Justification*. Cambridge University Press.

Manning, R. N. (1998). Functional explanation. In Craig (1998).

Mill, J. S. (1863). Utilitarianism. In *On Liberty and Utilitarianism*, pages 137–211. Bantam Books.

Millgram, E. and Thagard, P. (1996). Deliberative coherence. *Synthese*, 108(1):63–88.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press, 1995 edition.

Millikan, R. G. (1989). Biosemantics. In *White Queen Psychology and Other Essays for Alice*, pages 83–101. MIT Press.

O'Reilly, R. C. and Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press.

Papineau, D. (1993). *Philosophical Naturalism*. Blackwell Publishers.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press, 1987 edition.

Peirce, C. S. (1877). The fixation of belief. In Thayer (1982), pages 61–78.

Pinker, S. (1997). *How the Mind Works*. W. W. Norton & Company, 1999 edition.

Plantinga, A. (1993). *Warrant and Proper Function*. Oxford University Press.

Pollock, J. L. and Cruz, J. (1999). *Contemporary Theories of Knowledge*. Rowan & Littlefield Publishers, Inc., second edition.

Pryor, J. (2001). Highlights of recent epistemology. *The British Journal for the Philosophy of Science*, 52(1):95–124.

Putnam, H. (1975). The meaning of 'meaning'. In *Mind Language and Reality*, volume 2 of *Philosophical Papers*, pages 215–271. Cambridge University Press.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge University Press, 1995 edition.

Ram, A. and Leake, D. B., editors (1995). *Goal-Driven Learning*. MIT Press.

Rescher, N. (2001). *Cognitive Pragmatism*. University of Pittsburgh Press.

Rey, G. (2000). Meta-atheism. http://brindedcow.umd.edu/logo/meta-atheism.html.

Rorty, R. (1995). Is truth a goal of enquiry? Davidson vs. Wright. *Philosophical Quarterly*, 45(180):281–300.

Russell, B. (1907). The regressive method of discovering the premises of mathematics. In Lackey, D., editor, *Essays in Analysis*, chapter 13, pages 272–283. George Braziller.

Scott-Kakures, D. (2000). Motivated believing: Wishful and unwelcome. *Noûs*, 34(3):348–375.

Searle, J. R. (1983). *Intentionality*. Cambridge University Press, 1999 edition.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search: Adaptive Computation and Machine Learning*. MIT Press, second edition.

Sterelny, K. (1990). *The Representational Theory of Mind: An Introduction*. Blackwell Publishers.

Steup, M., editor (2001). *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. Oxford University Press.

Stich, S. (1990). *The Fragmentation of Reason*. MIT Press.

Stich, S. (1991a). *The Fragmentation of Reason*: Précis of two chapters. *Philosophy and Phenomenological Research*, 51(1):179–183.

Stich, S. (1991b). Evaluating cognitive strategies: A reply to Cohen, Goldman, Harman, and Lycan. *Philosophy and Phenomenological Research*, 51(1):207–213.

Thagard, P. (1988). *Computational Philosophy of Science*. MIT Press, 1993 edition.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton University Press, 1993 edition.

Thagard, P. (1999). *How Scientists Explain Disease*. Princeton University Press.

Thagard, P. (2000). *Coherence in Thought and Action*. MIT Press.

Thagard, P. (2001). How to make decisions: Coherence, emotion, and practical inference. In Millgram, E., editor, *Varieties of Practical Reasoning*, pages 355–371. MIT Press.

Thagard, P. (2002). How molecules matter to mental computation. *Philosophy of Science*, 69(3):429–446.

Thagard, P. and Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In Ram and Leake (1995), pages 439–454.

Thagard, P. and Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, 22(1):1–24.

Thayer, H. S., editor (1982). *Pragmatism: The Classic Writings*. Hackett Publishing Company.

Velleman, J. D. (2000). On the aim of belief. In *On the Possibility of Practical Reason*, pages 244–281. Oxford University Press.

Wedgwood, R. (2002). Internalism explained. *Philosophy and Phenomenological Research*, 65:349–369.

Wilford, J. N. (2001). New discoveries complicate the meaning of 'planet'. *The New York Times*, January 16.

ABSTRACT

BELIEF-DESIRE COHERENCE

by

Stephen Petersen

Chair: Eric Lormand

Broadly construed, this dissertation addresses a question central to normative epistemology: "what makes for good thinking?" My answer is a computational, internal, pragmatic, coherence epistemology. I call it, somewhat incompletely and inaccurately, "belief-desire coherence". It is designed to draw from (and contribute to) progress in artificial intelligence and cognitive psychology.

Probably the standard philosophical answer to "what makes for good thinking?" is a variation on "thinking directed toward the *truth*." I save the bulk of my arguments against this traditional alethic approach for the fourth chapter; in the first three, I motivate my positive, pragmatist alternative.

Chapter one focuses on the controversial topic of epistemic *guidance*. I begin by exploring what it is for a creature—natural or artificial—to be intelligent. Assuming that adaptability is fundamental to intelligence, and that learning is fundamental to adaptability, I develop an account of what it is for a creature to *learn* to think better. Starting with

a functional characterization of creatures, I argue that for a creature to learn better thinking requires a feedback mechanism internal to its cognition. The result is a naturalistic and more precise version of "internal" epistemology that captures and explains its basic intuitions. This characterization of internal epistemology suggests, in turn, that the internally available standard for better thinking is pragmatic, to do ultimately with fulfilling the creature's basic aims.

In the second chapter I consider the wishful thinking objection to any internal pragmatic epistemology, and in response argue for a modified coherentist approach to the evaluation of both beliefs and desires. An internally measurable standard of good thoughts, both desire-like and belief-like, is the level of coherence among them. The proposed coherence has foundational elements in the "default" thoughts that come with the fundamental design of the creature.

This pragmatic coherence measure, I claim, can provide the internal feedback required for learning. In the third chapter I show how to model this coherence and feedback computationally. Then, with the full theory in place, I outline its several advantages for cognitive science, accounts of folk psychology and emotions, and even ethics.