INFORMATION TO USERS

NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was microfilmed as received.

52

This reproduction is the best copy available.

UMI

ATTRIBUTIONS OF INFERENTIAL ERROR, EPISTEMIC VIRTUES, AND MODELS OF MINIMAL RATIONALITY

SEAN ALLEN-HERMANSON

A Thesis

in

The Department

of

Philosophy

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Arts of Philosophy at
Concordia University
Montreal, Quebec, Canada

June 1996

Canada

ABSTRACT

ATTRIBUTIONS OF INFERENTIAL ERROR, EPISTEMIC VIRTUES, AND MODELS OF MINIMAL RATIONALITY

SEAN ALLEN-HERMANSON

I defend an account of agency which holds that a more desirable, empirically accurate, *model of rationality* must reject certain well-entrenched assumptions about the connection between logic and good reasoning. *I argue following Cherniak, that models of rationality must use desire fulfilment, and not just logic or probability theory as the ultimate normative standard of rational choice.* Cherniak's theory is expanded upon, defended from both real and imagined *critics*, and *distinguished from the positions* of Stich, the Positivists and Relativists.

Acknowledgements

I gratefully acknowledge the kind support and encouragement of my supervisor, Professor Murray Clarke, during the preparation of this thesis. I would also like to extend thanks to the other members of my examination committee, professors Andrew Wayne, Jack Ornstein, and Sheila Mason for their helpful comments.

Table of Contents

List of Figures

Chapter One: A Defence of Minimal Rationality

Introduction

Recent work in psychology[1] suggests that most people are disposed to certain systematic inferential errors. It would seem that the manner in which people reason falls far short of theoretical ideals as defined by deductive logic and probability theory. Not surprisingly, the tone of the psychological literature on human inferential tendencies is strikingly pessimistic. Kornblith (1993, 70) cites psychologists Nisbett and Borgida for their remark that the experimental results bear "bleak implications for human rationality." In response, there has been a movement in the philosophical and empirical psychology communities to address this trend. What has emerged is a tendency to deny that there are necessarily bleak implications. Indeed, Kornblith goes so far as to say that the psychological research illuminates just how well our inferential systems actually perform! Others, such as Stich, come close to the view that virtually anything goes with respect to which inferences are to be counted as rational. Although it is doubtful there are grounds for Kornblith's optimism, the general approach of the philosophers is correct, I think. The psychologists' bleak implications only follow given certain, highly questionable, assumptions about what it is to be rational.

There is some agreement among several philosophers including Cherniak, Goldman, Kornblith, and Stich, that prescriptive epistemic norms should not be determined independent of psychological facts.[2] Unfortunately, some philosophers have focused laser-like on only those errors which can be easily explained away in various ways. Part of my intent in subsequent chapters will be to draw attention to those inferential errors which cannot be made consistent with various explanatory devices such as inferential heuristics,

speed-accuracy tradeoffs, and such.[3] Interestingly, there may be an analogy here with Quine's notion that "better translation imposes our logic". Quine's point was that we must assume that speakers of any newly discovered language accept the same rules of logic as we do if we want to translate the foreign language into something we find intelligible. Certain philosophers seem intent on interpreting any seemingly irrational choices made by the subjects of the psychologists' experiments in such a way as to construe those inferences as rational after all. The danger is that if we extend that strategy too far, and insist on finding ways to explain away *every* inferential error, we assume the view that nobody is really making irrational choices. If we are willing to assume that we do, in fact, make inferential errors, then such a strategy would result in the loss of the normative force integral to the concept of rationality. If our praise, condemnations, and suggestions about the way people *ought* to make decisions is in fact legitimate, then prescriptive epistemology must be preserved. Charitable interpretations of (so-called) inferential error must not be granted an unlimited scope. Kornblith seems to be particularly guilty of this sin.

If the rationalisation of inferential error is a sin, however, it is a sin of excess. Powerful arguments are at hand which down-play at least some of the psychologists' bleak implications. Cherniak, for instance, has emphasised that the "quick and dirty" inferential heuristics discovered by psychologists may sometimes actually be preferable to more formally correct systems. (Cherniak 1986a, 76) Certain facts about computational limitations and our physiology suggest that our inferential systems favour compromises between shortening computational speed and generating valid inferences. Another recurring theme in this thesis will be the exploration of the tension between constructing a realistic

2

descriptive epistemic model; a model, that is, determined by how we in fact do come to form beliefs and make decisions, and, preserving epistemic norms, or, the normative force which underpins prescriptive epistemology.

Following a standard convention, the use of double quote marks (" ") is meant for direct quotations and (rarely) as scare quotes. Single quote marks (' ') around words denote mention, and not use. For example: Beverley said, "I think I shall name the kitty 'Kitty'", and, Kitty is Beverley's "baby".

1.1 Background

The Standard Model of Rationality

For someone inclined towards a naturalistic world-view, a desirable Theory of Rationality must provide both an accurate description of actual inference-making as well as prescribe norms by which we could maximize our ability to gather and make use of knowledge. Cherniak's work suggests that Theories of Rationality are generally divisible into two broad categories, namely between those theories which presuppose highly idealised conceptions of rational agency and those which do not. Highly idealised models of rationality presume that rational agents must, in principle, obey all logical rules of inductive and deductive logic, all the time. The degree to which actual agents depart from these logical standards determines the extent to which they are making irrational choices. On the most extreme variants of this view the ascription of rational agency is not possible without assuming perfect adherence to the canons of inductive and deductive logic. The equivalence of rationality with logic is widely assumed in Economics, Political Science, Game-Theory,

3

.

Psychology, and Philosophy. It is, for example, implicitly assumed in the work of the Empiricists, Rationalists, Kant, and Chomsky,[4] and found more explicitly (and recently) in the work of Hintikka, Davidson, Dennett, Pollock, and Quine. All such Standard Models (SM) represent what shall sometimes be referred to as the 'Establishment View'. Given the copious amount of material on SM, a complete survey of every version is impossible here and unnecessary. Instead, I shall discuss certain defects common to any version of SM.

In contrast to the Establishment View, others have raised objections against both the descriptive power and prescriptive utility of SM. My attention shall be focused on Christopher Cherniak's criticisms of SM as well as the alternative he suggests in *Minimal Rationality*. Cherniak argues that the ideal requirement for agents to adhere to a universal logic should be relaxed. For Cherniak, a more accurate and normatively desirable model of rational agency need not, cannot, and ought not, conform to the epistemic requirements demanded by SM. Simply put, Cherniak's thesis is that agents should be required to follow only *some* rules of logic *some* of the time in order to meet the minimal requirements for rational agency. Cherniak's model of Minimal Rationality will usually be referred to as 'MR'.[5]

My intention is to examine Cherniak's proposal through a critical, but ultimately supportive perspective. This chapter will concentrate on establishing three primary goals: 1) explaining the differences between SM and MR, and; why MR is preferable over SM; 2) Responding to certain objections to MR; 3) Applying my discussion of MR to other areas of philosophy, particularly to Kripke's 'Quus-Plus' example and its critics. Later chapters will develop my examination of MR. In Chapter Two, I distinguish Cherniak's position from that

4

of Stich. Although, Stich sees his own views as closely aligned with Cherniak's, Stich ultimately draws conclusions quite at odds with Minimal Rationality. Cherniak, too, shall not emerge from Chapter Two unscathed. There I criticise him for advocating epistemic norms which are too weak. Cherniak incorrectly downplays the need for rational agents to maintain consistency between their beliefs. In the last chapter, I attempt to explain how Minimal Rationality can contribute to the Ameliorative Project in Epistemology' (AP). AP amounts to the task of improving our epistemic lives. I attempt to apply the insights of MR to the problem of improving the frequency of rational decision making.

Common Deficiencies of Standard Models

A variety of theories of rationality may be counted as representing the SM viewpoint, although idealised rationality conditions may be manifested within such theories in somewhat different ways. Since all SM theories equate rationality with formal rules, Cherniak handles their differences by arranging different rationality conditions on a continuum, ranging from the most idealised to progressively weaker and weaker standards. Here, maximally idealised theories maintain that agents must follow all logical inference rules, whereas theories with weaker standards require only some subset of those rules.

The condition of Deductive Closure is the strictest of the idealised rationality conditions he discusses. Under the Deductive Closure Condition agents are required to perform all sound inferences which follow from their belief set. Cherniak characterises Hintikka's (1962) as one such example of this view. Consider the claim that if p implies q and an agent believes p, then that agent *ought* to believe q. Hintikka "claims that it would be "indefensible" or "irrational" for someone to believe (or know) p and not to believe q

5

here." (Cherniak 1986a, 24) But a moments reflection shows that this cannot be right if it is meant that agents must draw *all* possible sound inferences. It would surely be irrational for someone to draw every trivial inference that followed from their belief-set. Creatures that wasted most of their time drawing uninteresting inferences like: if P then PvQ...if PvQ then PvQvR...and so on, would as Quine suggests, "have a pathetic but praiseworthy tendency to die before reproducing their kind." (Kornblith 1986, 5) The conclusion here is that even if it is granted that drawing sound inferences is important, the principle of Deductive Closure is much too strong to serve as a normative ideal. What is needed is some account as to when it is and is not appropriate to follow deductive closure.

The next idealised rationality condition considered by Cherniak is called the 'Ideal Inference Condition' (IIC) and represents an improvement over the extreme proposal of Deductive Closure. Under IIC agents must perform only those inferences which are apparently appropriate. Apparently appropriate inferences just are those that would fulfil some desire of the agent. For example, an inference which is required to fulfil the desire to survive would be one likely candidate for an apparently appropriate inference. Hence, under IIC agents would not be required to follow the trivial inferences that are implied by Deductive Closure.

Cherniak provides three reasons why IIC is still too strong to serve as a realistic rational ideal. First, although people tend to make mistakes we do not fail to attribute agent status to those who fail to make all the apparently appropriate inferences from their belief set. Suppose I turn the heat off in my apartment because I am going away for the weekend. However, I have carelessly forgotten that the batch of beer I had just brewed requires an

ambient air temperature of 20-25°C if the yeast is to work properly. My failure to infer that

I should have left the heat on in order to satisfy my desire for beer does not disqualify me

from being counted as a rational agent, (although I have made a careless mistake). Second,

ideal rationality conditions cannot be used to help predict human actions. This should be

considered a bad thing because we already know that predictions of human behaviour are

easy to make and extremely useful. Our ability to use folk-theoretic concepts ('beliefs',

'desires', 'hopes', 'fears', and so forth) enables us to interpret and anticipate the behaviours

of others: for instance, I can predict with great accuracy the reaction of my roommate if I

were to drink all of his Scotch, ceteris paribus.[6] But since idealised epistemic standards are

based on *logical* ideals we can never expect any actual agent to behave in perfect accordance

with such models. As Cherniak (1981) has pointed out, we cannot expect to find much

predictive efficacy from a theory which is never actually satisfied. Part of the problem may

be with the way in which the notion of possibility is framed within idealised theories of

rationality, (more on this point later). One might also pause to consider that if idealised

conditions were applicable to human agents, the deductive sciences would be trivial. But as

any instructor of an introductory course in deductive logic can attest, most undergraduates

do not find logic intuitive. The third problem with IIC is that it would actually be irrational

to try to draw all apparently appropriate inferences. The reason is that it may not be possible

for a creature to draw every inference that might be of some use to itself, for instance, as to

whether Goldbach's conjecture is true or not; some complex inferences are not possible for

finite agents to complete.[7] Also, it may well be the case that the costs of making many

otherwise appropriate inferences far outweigh the benefits to be obtained. This, of course,

.

7

is not to say that there is no relevance to the notion of apparently appropriate inferences

insofar as normative standards are concerned. The problem is that as a rationality condition,

it is still too strong to be useful. What is needed is to specify a subset of sound apparently

appropriate instances for a Minimal Normative Condition. That third condition concerns

the feasibility of an inference. In the next section I will explain Cherniak's notion of

feasibility and why he thinks that there is a principled basis for distinguishing between

feasible and unfeasible inferences

To summarise, Standard Models of Rationality share a common feature, namely, the

assumption that rationality amounts to following logical rules. Agents which are most

rational, on this view, just are those that follow deductive and inductive inference rules most

often. In contrast to this position Cherniak argues that rationality is not equivalent to

maximising one's adherence to logic, rather, rationality is best realised when agents follow

only some *subset* of logical rules. The character of that subset is defined by the notions of

soundness, appropriateness, and as shall be shown in the next section, feasibility. I call this

description of MR '*Weak*-Minimal Rationality'. This version is weak in the following sense:

since the only decision making procedures recognised as rational are those that are logically

valid, Weak-MR is not fundamentally distinct from the Establishment View. Both theories

share the view that logic *rules*, so to speak. Later on I discuss (and defend!) Cherniak's more

mature account of MR which I call '*Strong*-Minimal Rationality'. Strong-MR is

distinguished by the inclusion of certain non-formal judgements into the inventory of rational

decision procedures. Hence Strong-MR is much more radical a position than Weak-MR

because Strong-MR allows for the possibility that deductively invalid rules can be counted

as rational. Much more will be said later as to when formally invalid procedures can be counted as rational.

1.2 Minimal Agency: Motivations and Exposition

Complexity Theory and the Notion of Feasibility

As has been previously mentioned, one way of putting Cherniak's thesis of Minimal Agency (the Weak version) can be summed up with the slogan:

> In order to be minimally rational, an agent must follow *at least some* of the rules of logical inference.

The phrase 'at least some' is emphasised because it represents the conceptual shift away from the universal quantifier inherent to idealised theories of rationality. Later I shall discuss more precise explications of the existential quantifier in greater detail. Cherniak's ideas are, in part, motivated by certain developments in the field of complexity theory. In Chapter Four of *Minimal Agency*, Cherniak identifies certain relationships between computational complexity theory, the psychological literature on inferential error, and the Minimal Agency model of rationality.

According to Cherniak, complexity theory provides a principled basis for distinguishing between practically feasible and unfeasible computational tasks. In complexity theory, an algorithm is considered feasible if its execution time grows as a polynomial function of the size of input instances of the problem. An algorithm is unfeasible (computationally intractable), if computational time grows faster than this, usually as an exponential function. In other words, even on the most advanced supercomputer, some algorithms are not feasible to compute because they require exponential time to complete.

9

To illustrate, Cherniak gives the example a theorem of only 617 symbols. To prove such a theorem "would require a network with so many boolean elements that, even if each were the size of a proton (with infinitely thin interconnecting wires), the machine would exceed the volume of the entire known universe." (Cherniak 1986a, 90)[8]

A partial solution to the problem of computational intractability within complexity theory is supplied through Rabin's notion of a probabilistic proof. Some kinds of mathematical proofs are so complex as to be practically impossible to complete. Rabin devised a method by which the requirement for certainty is relaxed. In this way otherwise intractable problems can be solved although there is no guarantee of truth. Note, however, that the estimates of error can be very precise; Cherniak cites one example in which the probability of error is one in a billion. (Cherniak 1986a, 90) Cherniak sees Rabin's general strategy as a desirable alternative to the Cartesian-like project of seeking absolute certainty since in some cases the demand for apodetic certainty must lead to paralysis. Similarly, human beings are no less free from the constraints imposed by computational intractability. Our lives are short, our memories finite, and we are generally error prone. Cherniak concludes that a kind of inferential speed-accuracy tradeoff is necessary for minimal rationality. The certainty of the correctness of our inferences is traded in return for quicker response times, surely an advantage in a fast paced Darwinian world.

Cherniak theorises that some middle ground between pure guesswork, the quickest and dirtiest procedure, and a reliable but uneconomically slow decision procedure best explains our inferential tendencies. Sometimes a quick answer we are unsure of is preferable to a process which is more reliable but slower. It is easy to think of examples where such

10

a preference becomes a necessity. If I am playing a game of chess under a time limit I do not have the luxury to work out the implications of every possible move. A rule of thumb which provides near optimal moves very quickly would give me an advantage over a player with a perfect but agonisingly slow decision making procedure.

Cannot Implies Ought Not

In H.P. Lovecraft's "The Dream Quest of Unknown Kadath", Randolph Carter's sense of ethical obligation runs up against the physical limitations imposed by the situation:

> The number of malodorous moonbeasts about that greenish fire was very great, and Carter saw that he could do nothing now to save his former allies...For a moment he pondered on what he ought to do...

Like Carter, our normative lives are constrained by the physical limits that come with being human. Unlike Carter, however, we cannot hope for a lucky improvement of the situation. Barring some miraculous mutation or advance in bioengineering, it seems our cognitive deficiencies are here to stay; this is, as Cherniak puts it, our *finitary predicament*. Given this, it can be understood how Cherniak's basic argument for MR is, in a sense, Kantian: If ought implies can, then it follows by *modus tollens*, that *cannot implies ought not*: $A \rightarrow B :: \sim B \rightarrow \sim A$. Cherniak thinks our epistemic norms should be greatly constrained in light of our limitations as finite beings. For instance, our brains have a limited capacity to store and retrieve memories. Furthermore, our memories are divided into a small short-term active memory and a larger but relatively inactive long-term memory. We have finite life spans, and are prone to making mistakes of all kinds. These and other failings conspire to ensure that certain inferences will always be beyond our abilities. For instance, I would die before I could complete an inference which required one-thousand years of computational time.

Also, as has been pointed out earlier, just because an inference is sound does not mean it would be reasonable to make it. Cherniak remarks that "a person could waste his entire lifetime, probably a short one, making only such uninteresting inferences." (Cherniak 1986a, 24) Inferences that are interesting are those that are be beneficial with respect to the creature's survival, or, if you prefer, those that fulfil the creature's desires. Indeed, Cherniak is right to claim that "*not* making the vast majority of sound and feasible inferences is not irrational; it is rational." (Cherniak 1986a, 24) The subset of inferences that are sound, appropriate, and feasible comprise the Weak (in the sense mentioned earlier) Minimal Rationality Normative Condition. In section 1.3 I shall discuss the composition of the inference set that makes up the Minimal Normative Condition in more detail.

The Interpretation of 'Possible'

Possible world semantics might be helpful to further distinguish MR from SM. The interpretation of 'possibility' as in, 'it is possible for any agent to make inference _____.' according to SM means 'logical possibility'. I interpret Cherniak's argument as proposing that we need to distinguish that which is logically possible from that which is only practically possible. There is a correlation between the relative positions of the various rationality conditions Cherniak describes and these various interpretations of possibility. Consider again the idea that we ought not do what we cannot do. If I mean 'cannot' in the sense of logical impossibility then I mean 'ought not' in that same sense. 'Practically possible' here simply means a subset of what is physically possible. Normative requirements to follow (say) deductive closure would seem to ignore these different metaphysical senses of 'possibility' If it is physically impossible for me to do something, whether that something is saving my

allies, following closure, or whatever, then there can be no obligation to do so (at least in *this* possible world).

Normative and Descriptive MR

Minimal Normative Rationality must also be distinguished from Minimal Descriptive Rationality. As has been mentioned, Minimal Normative Rationality (the Weak version) consists of three parts: 1) The soundness of the inference, 2) Its feasibility, and, 3) Its apparent usefulness according to the agent's belief-desire set. Minimal Descriptive Rationality, however, is itself a subset of the subset of logical rules defined by the normative conditions. Since we must expect people to make mistakes, the minimal requirement for ascription of rational agent status must allow for performance worse than the Minimal Normative Ideal.[9] The results of these considerations are represented in figure1 below:

Fig. 1: Schema of Various Rationality Conditions

All sound inferences.. Maximum Normative Rational Ideal
(Condition of Deductive Closure)   (Logically Possible)
                  *
                  *
                  *
                  *

All apparently-appropriate sound inferences...... Belief-Desire Model Ideal
(Ideal Inference Condition)
                  *
                  *
                  *

All appropriate sound inferences......................... Weak Minimal Rationality Normative Ideal
that are feasible.   (Physically Possible)
                  *
                  *
                  *
                  *

Some appropriate sound inferences..................... Weak Minimal Rationality Descriptive
(where 'some' approaches 'all')   Limit
                  *   (Becoming less probable)

13

```
*
*
*.....................................................Average performance of most actual agents
*                                                     (Most probable)
*
*
*
*
Some appropriate sound inferences..................... Weak Minimal Rationality  Lower
(where 'some' approaches the                           Descriptive Limit
minimum required for agency)                           (Becoming less probable)
                  *
                  *
Fewer inferences than required
by the Minimal Lower Descriptive Limit........... Not rational
```

The distinction between Minimal Normative and Minimal Descriptive Rationality explains

why Cherniak's position is not compatible with what Kornblith calls 'Psychologism', that

is, the view that the way we arrive at our beliefs just is the way we ought to, or, somewhat

less charitably "the organism is always right." (Cherniak 1986, 25) Minimal Deductive

ability is weaker than the Minimal Ideal because we must expect a Minimal Agent's actual

behaviour to fall far short of what it should be. Cherniak's Normative Ideal lies at the

extreme edge of physical possibility so although an agent could conceivably live up to the

Minimal Ideal in the actual world, this is thought to be unlikely. So ascription of Minimal

Agency only depends on agents adhering to some subset of the requirements of Minimal

Normative Rationality. It is important to keep in mind, though, that mere ascription of

rationality does not constitute a normative standard, hence the need for the Minimal

Normative Ideal remains. It should also be noted that MR is compatible with a weaker form

of Psychologism known as 'Ballpark Psychologism', or, the view that the way we reason is

roughly like the way we ought to.

14

Strong Minimal Rationality

Before proceeding further, it is necessary to point out a tension between Cherniak's explication of the conditions for Minimal Agency in chapter one and chapter four of *Minimal Rationality*. The thesis of chapter four is that formally incorrect decision making procedures may be preferable given certain feasibility constraints, particularly with respect to computational time. Kornblith (1992), for instance, has provided one explanation in which a formal error may underlie a heuristic useful in picking out natural kinds. The trouble is that Cherniak's chapter one Minimal Rationality Conditions (soundness; appropriateness; feasibility) would seem to rule out including any formally incorrect procedures. Since the only allowable inferences are a subset of all *sound* rules, any rule which is unsound cannot be included. Up until now I have focused my attention on this first version, 'Weak Minimal Rationality'. Fortunately, MR can be easily corrected to account for formally invalid heuristics. In order to allow for formally incorrect procedures of the kind described in Cherniak's chapter four, the requirement for soundness should be dropped altogether. Although it is possible that many or most of a Minimal Agent's inferences will happen to be sound, this should not be an a priori requirement already built into the Minimal Normative Rationality Conditions. Any unsound inferences that are included must still fall within the subset of all apparently appropriate and feasible inferences. It should be noted that Cherniak seems to be unaware of the differences between the two versions of MR found in his book. Indeed, he seems to think there is only one theory being offered. It is for this reason I have made the distinction between Strong and Weak Minimal Rationality. Henceforth, unless stated otherwise, let it be assumed that 'MR' or 'Minimal Rationality' now refers to Strong-

MR. Strong MR just is Weak MR without the requirement for soundness.

1.3 Vagueness and the Explication of 'some'.

The Objection

Goldman (1992) has criticized Cherniak for the vagueness of the purely formal characterisation of the quantifier 'some' as in "A minimal agent must make *some* sound inferences from her belief set." Goldman is correct insofar as A) Cherniak is in fact vague on this point, and, B) the issue is too important not to address directly. There is a world of difference between the notions of almost none and almost all. I would not be overly concerned to learn that there is almost no Mercury floating around in my blood. In contrast, I would be upset if 60ccs of that heavy metal was injected into my arm. Clearly, when Cherniak says "some" he means more than "just a few." Given the huge numbers of inferences anyone typically makes, it seems fair to think that the number of correct inferences required to fulfil the Minimal Descriptive Condition would be quite large, (say) around at least fifty-percent. Still it is of interest to further explain what 'some' amounts to. Yet Goldman dismisses Cherniak's account without considering how 'some' could be explicated beyond a purely formal characterisation. Goldman also mistakenly seems to think that this vagueness results in a theory with no predictive content and therefore little scientific interest.

Goldman's objections are presented as brief remarks and lack support. Regarding the predictive efficacy of MR, I think Goldman's criticism is misplaced. Whether or not MR has predictive power should be of secondary interest, at best. MR is meant as a *philosophical* theory, and, as such, it matters little if it is or is not of any scientific interest. If MR is a

scientific theory, then it must be the source of some novel predictions. I confess that it would not seem to offer any. Even so, as a theory of rationality, MR has much to offer. First, MR offers to give an account of rationality more in step with the facts than the Establishment View's alternative. Second, MR shall be used to generate rational norms. Assuming we reject psychologism, science is insufficient to provide us with epistemic norms and prescriptions.

MR does predict that different agents could have radically different inference feasibility orderings and this, at least, does seem to be a matter that is empirically decidable. The notion of 'feasibility orderings' is better explained within the background of Cherniak's argument against what is known as the 'Fixed Bridgehead' view of rationality.

Stich (1990, 40) mentions Hollis as one adherent of the position that any rational agent must manifest a particular subset of logical inference rules; a bridgehead, as it were, of true and rational beliefs and inferences. To support this view, Hollis supposes that most people find some particular kinds of inferences trivial and others quite hard; we can thus imagine a sort of feasibility ordering of inferences from those we tend to find trivial to those we find counter-intuitive. In reply, Cherniak asks us to imagine a creature whose feasibility ordering is inverted with respect to that of a typical human. If such an agent could exist then the Fixed Bridgehead view must be mistaken. In contrast, on MR there is no particular rule that any agent must follow. Cherniak's thought experiment seems to be at least partially confirmed by certain anthropological evidence which I discuss in chapter three. One subject, a Kashdan peasant, would seem to resemble an agent with a radically different inference feasibility ordering than that of a typical North American in some respects. It is crucial to

17

notice, though, that the Kahsdan's feasibility ordering is not truly inverted with respect to our own. For example, like us the Kashdan finds certain simple syllogisms easy. Since nobody has been found whose ordering is inverted with respect to (say) that of a typical North American, it is unclear whether, agent feasibility orderings really can differ as greatly as Cherniak supposes.

The issue of vagueness Goldman raises is echoed by Hooker's similar concerns about Cherniak's lack of an extensive formulation of 'at least some'. (Hooker 1994, 216; also Footnote 27) The Minimal Rationality requires making a conceptual shift away from universal logical inferential requirements for agents. Having a clear, well defined, notion of what the scope of the Minimal Normative requirements are should be of great interest to any proponent of Minimal Agency. By giving a more detailed explication of 'some', Cherniak's theory is at once expanded upon, clarified, and defended.

The Reply

To begin, one might be tempted to think that a good way to respond to charges of vagueness would consist in giving a particular subset of inferential rules for agents to follow, or alternately, it may mean giving a certain minimal number of rules to follow, regardless of what those rules are. Both of these strategies are undesirable, as I shall explain.

Here it is important not to misconstrue Cherniak's proposal. Goldman seems to be asking for a specific set of rules, or minimal number of rules, that any agent must follow. But such a project is not MR. Goldman's objection can be levelled against certain other views, notably, the Fixed Bridgehead position. This view is open to Goldman's objection since Hollis is obscure as to which specific rules must be present in any agent. Cherniak, on

18

the other hand, makes no such claim. On MR there simply is no rule anyone must follow. Ascriptions of rationality are tentative and more or less ambiguous depending on how many inferential errors an agent makes. Hence it is said with great confidence that a candidate who never makes a mistake is rational, and with lesser degrees of confidence as the incidence of error increases.

To give a list of specific rules or minimum numbers of rules would amount to giving necessary and sufficient conditions for what it is to be rational and as has been often pointed out before, this is not a plausible enterprise.[10] The conceptual analysis of 'rationality' is a project fraught with extreme difficulties. Putnam, for instance, has remarked that it is unlikely that we could find universal generalisations obeyed by all instances of rational belief. (Putnam 1981 chapter 5) Black (1986) has also made much of the context sensitivity of rationality. Consider:

Is it Rational...

1. To presume that what is written in your local newspaper is true?

2. To pay $20 for a one-in-a-thousand chance of winning $100,000?

3. To walk home alone at night?

4. To think that a valid argument with true premises has a true conclusion?

Black's point is that there is no consensus among either lay persons or experts with respect to the correct usage of 'rational' and 'irrational'. Obviously, the question of whether or not items 1 and 3 are rational simpliciter is nonsensical. Questions such as "Am I male or female?", and, "Do I live in an area with a high crime rate?", and, "Do I live in Sarajevo or Halifax?", and so on must be answered in order to fix the relevant background before one

can sensibly ask the question "Is it rational *for me* to walk home at night?" It might be objected that I have rigged these examples by focusing only on questions that are obviously highly contextual, but with a bit of imagination it is easy to see that many examples which, at a glance, look context-independent are in fact not. With that purpose in mind, consider item 2. If you are a Bayesian, you would consider that action reasonable: the utility expectation of the bet is far greater than the risk involved. But suppose you are starving *today* and your last $20 is needed for food, or alternatively, suppose you have made a promise not to gamble. Of course, it may be that there are some context-independent principles of rationality, perhaps something like item 4, or the principle of utility maximisation. But even if it is true that there are a few such principles, this would hardly be enough to show that a comprehensive definition of 'some' is a plausible idea. That, it seems, is not likely unless *all* correct usages of 'rationality' can be specified.

Necessary Vagueness

Although it would be a good thing to dissipate some of the vagueness surrounding the existential quantifier, a certain amount of opacity is actually desirable. This is not in itself a bad thing. Just because I cannot give all of the necessary and sufficient conditions for the extension of the word 'pig', it hardly follows that there are no pigs or that there are not knowable correct and incorrect usages of 'pig'.[11] A failure to produce a list of necessary and sufficient conditions for 'rationality' should similarly imply no dire consequences. Consider also that Cherniak's Minimal Normative Condition identifies a number of apparently appropriate inferences a creature should make, and, as such, "any one or more of these properties can be absent, and yet the creature may still qualify as having a cognitive

system." (Cherniak 1986a, 19) Certainly rationality cannot be preserved if all of the properties are absent, but the distinction between what Cherniak calls the cognitive "vanishing point" and the Minimal Normative Conditions is not sharp . Yet, "a gray area does not imply no distinction between black and white." (Cherniak 1986a, 19) Quite so. The gradual transition from day to night hardly implies that daytime and nighttime are not discrete states. Hence what kind of evidence will count towards indication of a cognitive system is also highly context-sensitive.

Cherniak would agree that there are no pat answers: "One may wonder whether being minimally rational is not like being a little bit pregnant- an illegitimate notion..." (Cherniak 1986a, 19) But the Goldman-type worry that the Minimal Rationality Conditions must be empty of content is unfounded. As Cherniak points out, the blurred boundary of physical objects from the perspective of micro-physics does not prevent us from navigating the world from our everyday macro-perspective. Again, as with the 'pig' example, the vagueness of a concept need not preclude its use nor imply that its referent does not exist. On the other hand, the vagueness of a concept does not automatically imply that there are any clearly defined cases either. Yet with respect to MR there will be clear cases, clear that is, in the sense that we can provide at least some unambiguous judgements about a potential agent's claim to rational agency. At the extreme lower end of the scale is a case in which a candidate for agency makes only random inferences. Something cannot be counted as rational if it makes no, or only random, inferences. An "agent" which never makes any inferences is not an agent at all. At the other extreme, the clearest case of rational agency is one in which an agent always (or almost always) makes choices which fulfil her desires. This is so even in

spite of the ambiguity in the phrase 'fulfil her desires' here. Well known counter-examples to the (simplified) desire-satisfaction model of rationality I endorse often depend on examples which emphasise how a single agent can hold multiple, conflicting, desires. This point, however, has no bearing on whether or not there are clear cases of rationality and irrationality. Clearly, one can *suppose* there is an agent who happens to have desires that do not conflict. If such an agent did in fact always or almost always act in ways that satisfied her desires, she would rightly be counted as optimally rational. Although most people would fall somewhere in between these two extreme examples, the conceptual extremes show that there are clear cases on either side of the continuum of rational agency.

Consider also how the difficulty of resolving the issue of vagueness may depend somewhat on whether it is an *agent* being evaluated, or simply a particular *action* of an agent. The truth is that just about anyone will be counted as rational, in the sense of having beliefs, desires, and other intentional states even if they only rarely make rational inferences. No one likes to be called 'irrational' and certainly everyone has behaved in an irrational manner at some time. So when we use 'irrational' in the sense of evaluating a whole person, the activity is non-specific. Yet, evaluation can be much more specific when restricted to particular choices and behaviours. Particular choices, for example, can be judged based on how well that inference performed within that agent's own belief-desire set; within her own set-of-rules, so to speak. In contrast, simply calling someone 'irrational' or 'rational' is often a vague and unconstructive activity. Until these kinds of domain specific questions are posed and answered most actions cannot be evaluated with respect to rationality. Again, once the background is fixed, desire fulfilment becomes the paramount aspect of evaluation. Here,

then, adherence to logic need not be perfect adherence.

Turing's Test

In our ascriptions of rationality there may also be an analogy to the Turing Test or 'Imitation Game'. If the Turing Test is right, we ascribe rationality by being convinced by a speaker's activities that they are conscious. It is sometimes objected that the Turing Test is just a form of behaviourism, after all, it is conceivable that a sophisticated but unthinking machine could pass it. Yet, it seems that "we treat *each other* as black boxes, relying on our observation of apparently intelligent behaviour to ground our belief in other minds." (Dennett 1988, 94) Not applying Turing Test-like games leads to a kind of solipsism: if one rejects the outward behaviour of others as evidence for mentation, what other reason is there to believe there are other minds? Let it be supposed that if some computational system is sophisticated enough to pass the Turing test, we should adopt the intentional stance towards it; "my attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul." (Wittgenstein 1993, 178)

Even if it is true that only a thinking thing could pass the Turing Test, however, it is also conceivable that a thinking thing could fail to pass it. This point shows why the Turing Test is distinct from behaviourist assumptions: Mentalistic notions still play a role in explaining certain failures of thinking things to pass the Turing Test. Real agents may perform badly for any number of reasons, and indeed, the behaviour of actual agents is often a misindicator of their actual mental states. If I am in pain and I am asked how I am feeling, I could lie or otherwise conceal my actual condition for many reasons. Perhaps I am a playing some kind of joke, or I am a Stoic, or whatever. So the Test is not perfect. If

23

something does not pass the Turing Test it may or may not be conscious. Still it might be useful to imagine a kind of Rationality-Test. A Rationality-Test, like Turing's, would depend on subjects passing a minimal amount of inferences. In fact, such a Rationality-Test may be just another way of describing the Turing Test. Do not the notions of consciousness and rationality go together?

Of course, although rationality implies consciousness, consciousness need not imply rationality: after all, real agents often act irrationally (example: my family). But a Rationality-Test analogous to the Turing Test may still be useful. Notice that the Turing test is highly contextual; it is implausible that one could specify exactly the necessary questions and answers needed to pass the Turing-test. The point is that this in no way makes Turing's Test an arbitrary matter. Indeed, one expects that a successful candidate must meet quite stringent criteria. In the same way, you cannot give exact numbers of rules or specific rules that any rational agent must follow. So the content of 'some' as in "a rational agent must follow some of the rules of logic" gets fixed in a similar kind of way as the questions and answers to the Turing Test get fixed: it is partly a matter of convincing the community of rational agents that you are one of them.

Concluding Remarks Regarding 'Some'

This concern to recognize various aspects of context with respect to rationality is stressed repeatedly in Hooker (1994). He points out that the context-dependency of costs and benefits in rational action leads to the context-dependence of rationality for finite agents. The costs of being rational in one respect may outweigh the benefits to be obtained. So, it is irrational to stand around drawing trivial inferences from one's sense-data while an angry

lion charges towards you. But "the idealisations of reason, by contrast, are context-free-they call for global consistency, period." (Hooker 1994, 196)

Hooker, too, thinks that 'some' cannot be spelled out in advance of extensive empirical investigation. It partly depends on details of our cognitive psychology. Also, the importance and success of attempts to expand on 'some' depend on the context of the attempt, and the direct and opportunity costs and benefits (by context). Hooker claims "it is neither possible nor desirable to spell out the 'at least some' in a context free manner, so one must embed the project in a full account of human society." (Hooker 1994, note27) The reason why this is so is because there are several factors inherent to rationality that cannot be accounted for on the idealised models. Those features are context-dependence, resource allocation (in the sense of costs and benefits), risk, heuristic inference strategies and non-formal judgements. In the next section I shall explain why these factors presuppose a nonidealised conception of reason as well as why we should think these things are inherent to a complete account of rationality.

1.4 Idealisations

Theoretical Idealisations and Minimal Rationality.

The issue of theoretical idealisations with respect to theories of rationality figures prominently in the examination of Minimal Rationality. In this section, I shall examine recent discussion of Kripke's *Wittgenstein on Rules and Private Languages* which uncovers a dispute over the issue of theoretical idealisations in general. According to Kripke, idealisations with respect to the mind are unjustified, yet his critics argue that he is in fact

unjustified in making a distinction between theories of mind and other scientific theories. I shall use the discussion of Kripke's 'Quus-Plus' example as a framework through which the issue of theoretical idealisations about cognition is illuminated. The issue of idealisation is particularly salient in Quus-Plus and I intend for discussion of that issue to shed light on the issue of idealisation in regards to Cherniak's theory.

Quus-Plus

Before proceeding with the discussion of idealisations, some explanation of what Quus-Plus is all about is in order. Kripke develops his version of what he calls 'The Wittgenstein Paradox'. In Section 201 of *Philosophical Investigations*, Wittgenstein writes, "this was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule." For Kripke, the Wittgenstein Paradox is a new form of philosophical scepticism.[12] Kripke's variant is developed as a mathematical example. Consider first the function x plus y = z. Kripke asks us to suppose that there is a function 'quus' such that: quus = plus if x, y < 57, otherwise x quus y = 5. Suppose now that I have never before made the computation 58+67. The sceptic poses the challenge in two ways: first, the sceptic denies there is any difference of fact between the interpretation that I am following quus or plus. Second, the sceptic questions whether I have any reason to think I should answer '125' rather than '5'. Kripke's sceptic doubts whether any instruction I have given myself in the past compels me to think I followed plus and not quus: "Perhaps when I used the term 'plus' in the *past*, I always meant quus..." (Kripke 1982, 13) So a reply to the sceptic must identify what difference of fact shows that I did mean plus and not quus, as well as show why I am justified in answering '125', not '5'.

26

An initial reply to Kripke might argue something akin to the following. Supposing that we can make certain abstractions about our rule-following behaviour, that is, about memory limitations, lifespan, slips of the pen, and so on, could it not be that we can say something like, "there is a psychological generalisation such that, *ceteris paribus* when one is presented with sums one will follow plus and not quus?" Thus it has been argued that the burden of proof lies on Kripke to explain why idealisations about the mind are unjustified.

Pietroski and Rey (1994) defend idealisations about the mind as a part of their defence of a *ceteris paribus* account of natural laws. Relevant to this discussion is their complaint that idealisations about the mind have been singled out by Kripke as illegitimate for no good reason. They discuss Kripke's claim that such idealisations are implausible. Kripke argues that we have no idea what would happen "if my brain were stuffed with extra brain matter, or if my life were prolonged by some magic elixir." He concludes, "such speculation should be left to science fiction writers and futurologists." (Kripke 1982, 27) Regarding Kripke, Pietroski and Rey write:

> Perhaps there is something special about idealizations about the mind. But Kripke offers no reason for thinking that they are somehow less legitimate than idealizations about any other region of the world. Abstraction from mortality, brain size, or memory limitations would seem no worse off than abstraction from elasticity, electric charge, or some of the effects of the Big Bang. (Pietroski et al. 1995, 107)

They argue that Kripke's reservations against idealisations about the mind lead to an absurdity, namely, if his arguments were applied elsewhere "most-if not all-of genuine science would have to be left to science fiction!" (Pietroski and Rey 1995, 107) I believe that arguments of this kind can be reformulated into an interesting and worthy objection to

Cherniak's project. Nevertheless, there are replies to such objections at hand. Although Kripke's claims lack support, they are well motivated. I shall argue that Hooker's distinction between degenerate and simplifying theoretical idealisations provides the justification for rejecting certain idealisations about the mind.

The Argument From Scientific Idealisations

A fundamental premise behind Minimal Rationality is the rejection of the kind of theoretical idealisations made within Standard Models of rational agency. MR rejects the notion that rational agents should follow Deductive Closure, for instance. Yet as the preceding discussion of quus-plus has shown, the legitimate use of idealisations in scientific theories is commonplace. Given this, a critic of MR might be motivated to raise the following objection: could it not be argued that if we are justified in making idealisations in scientific theories of various kinds, we are similarly justified in making idealisations in our theories of mental phenomenon? In this section, I shall explore and address this objection. Through arguments given by Hooker (1994) I hope to show why idealisations about the mind are unwarranted thus lending more initial plausibility to Minimal Rationality. After this task is completed, the resolution of the Scientific Idealisations argument with respect to Cherniak will be redeployed as a rejoinder to critics of Kripke.

Pietroski and Rey's solution to Kripke's puzzle leans on the notion of abstracting away from the finite limitations of human memory and lifespan. If such an approach is, in fact, correct, it might be applied beyond Kripke as a general objection against all Minimal Rationality-type models of rational agency. The Argument from Scientific Idealisations (ASI) might go something like this:

28

P1. Idealisations occur legitimately in all scientific theories.
P2. Desirable theories about the mind are scientific theories.
C1. It is justifiable to make idealisations about the mind.

If correct, this argument would do much to discredit Cherniak's project. Let me put the objection this way: if Pietroski et al. are right, and abstractions about memory capacity or life span are no different in kind than abstractions about charge, frictionless surfaces, elliptical orbits, and such then it would seem that the motivations behind Minimal Rationality are deeply misguided. It would be as if someone proposed we replace the ideal gas law with a 'minimal' gas law which included enough provisoes to account for every conceivable perturbation and idiosyncrasy of any particular gas molecule. It would be highly surprising to find that an object was found to fall at exactly 9.8 m/s^2. And even after we take other factors into account, such as air resistance, errors of measurement could still be counted on to produce observations which deviated slightly from expected behaviour. So, it would be silly to suggest that even rough approximations such as $PV=RT$ are empirically barren on the grounds that no actual molecules ever behave exactly as that idealised version of the gas law predicts. It might be thought then that an analogous argument could be made with respect to theories about minimal rationality. Indeed, one might go so far as to say that Cherniak's criticism of SM implies just such an absurdity. Recall Cherniak's point that no agent could ever behave in ways that SM actually predicts. Does this also imply that Cherniak should also regard the gas law *mutatis mutandis* as equally devoid of empirical content? Perhaps Cherniak is in trouble.

The merits of this objection, I think, should not be overestimated. In reply I argue that ASI, and the implications drawn from it regarding MR, gloss over the very crucial fact

that just because an idealisation *occurs* in a scientific theory does not make it *justified*. Thanks to Hooker (1994), it can be shown that there are at least two kinds of idealisations made within scientific theories, only one of which is made legitimately. If so, it seems that premise one of ASI can be undercut. Indeed, in the next section I argue that the onus is on any advocate of ASI to explain why we should a priori think that just because an idealisation is made as a part of a scientific theory it is automatically legitimate. Scientific theories can be refuted and, as such, particular aspects of particular scientific theories should not necessarily be taken as the final word on good theory construction.

Hooker's Distinction

Pietroski and Rey are correct in their criticism of Kripke insofar as the burden of proof should be on him to explain what makes idealisations about the mind so different than in other areas of science. In order to meet this challenge what is needed are reasons to think that idealisations about the mind are misleading, flawed, or otherwise unjustified. With respect to his own comments, Kripke fails to do this. Fortunately for Cherniak and his followers the resources needed to distinguish between reasonable and unreasonable idealisations are found in Hooker (1994). Hooker introduces a distinction between what he calls 'simplifying' and 'degenerate' idealisations within *scientific* theories. I stress the word 'scientific' in order to emphasise that there is nothing particularly idiosyncratic about rejecting idealisations about the mind. In short, I challenge the truth of premise one of ASI. Theoretical idealisations are not always legitimate just because they occur in other accepted scientific theories.

It is Hooker's contention that there are two distinct kinds of idealisations.[13] The first

kind, *simplifying* idealisations, represent correctable simplifications in which a relatively non-idealised theory is approximated by an idealised one. For example, the Newtonian law of free fall, $F=mg$ is an idealised approximation of the behaviour of falling bodies. The description of gravitational acceleration is made more correct and more complex when a the effects of air resistance are added. Similarly, the ideal gas law is corrected when gas molecules are considered to have a finite volume. In both cases the idealisations represent a reversible departure from empirical behaviour. The theoretical deviation from empirical behaviour is corrected when additional correcting terms or provisoes are added. Hooker calls such deviations 'quantitative' departures from actual behaviour. Hooker points out that virtually all scientific theories make use of idealisations of this kind.

The second kind of idealisations are *degenerate* in the sense that they are not just quantitative departures from observed behaviour, but also represent a *conceptual inaccuracy*. One instance Hooker gives of a degenerate idealisation is the approximation of the speed of light with infinite speed. Granted, under many circumstances such an equivalence produces no ill effects. Under most circumstances Newtonian mechanics may be as useful as Relativistic physics, often even more so. Nevertheless, we now know that it would be a misrepresentation of the structure of space-time to assume that the speed of light is infinite. In order to put forward relativistic theories of space-time it was necessary to specify the non-infinite speed of light: c. Additional terms, provisoes, or corrections cannot capture relativistic space-time as a complication of Newtonian space-time. The main difference between simplifying and degenerate idealisations is that an idealisation is a simplification when actual behaviour can be represented as a correctable deviation from idealised

31

behaviour.

Reply to ASI

With this distinction in mind, the degeneracy of idealisations about the mind can be established. Hooker sums up the intent here nicely:

> Cherniak's arguments are most fruitfully understood as supporting the thesis that analytic theory of rational agents is a degenerate idealisation, not a simplifying idealisation, of an adequate theory of finite rational agents. (Hooker 1994, 208)

Yet one may wonder why should we regard idealisations about the mind as degenerate in the first place? The reason why idealisations about the mind are degenerate is because they are not correctable simplifications of actual rational behaviour. Of course, in order to avoid circularity, it is necessary to explain why we should think idealisations about the mind are not correctable simplifications of observed behaviour. The various rationality conditions which fall under SM necessarily fail to account for several important features inherent to rationality. Those features include, context-dependence, resource allocation (in terms of costs and benefits), risk, heuristics, and non-formal judgement (Hooker 1994, 210) Each of these notions are inherent to rationality, hence any theory which cannot account for them must be deficient. Just as Newtonian physics represents a conceptual distortion of space-time, so too do overly idealised models of agency distort our picture of rational agency.

If idealisations about rationality are degenerate because they cannot account for these five features, then it must be that these features are *fundamental* to rationality. Why should we think they are? Allow me to explain. The first feature refers to the costs and benefits of being rational in one respect rather than another. In terms of resources (cognitive and otherwise) any given rational organism must strike a balance between the costs of making

32

a particular inference, pursuing a line of inquiry, and so forth, and the benefits to be obtained. Naturally the costs and benefits associated with most cases of thinking will be highly sensitive to the context of the situation. For instance, the benefits of following certain rules-of-thumb, such as "If danger then climb tree" can work quite well in general, but may ultimately prove to be more costly than beneficial supposing one is being chased by a bear. These kinds of considerations make it plain that context-dependence is firmly rooted in the determination of costs and benefits and in turn to the notion of rationality itself.

The second features, risk and heuristics, are closely related. Because of the changing character of costs and benefits from context to context and the limitations of human inferential capacities, it seems that we are required to follow informal or less than ideal rational procedures. Such so-called 'heuristics', then, necessarily involve a certain risk: since a heuristic is an imperfect strategy, we must expect that sometimes we will end up making erroneous decisions. Inferences which depend on quick and dirty heuristics are especially risky. Even so, it is likely that we are often faced with the choice of making a decision based on that minimal kind of heuristic and making no decision at all. In such a case we might fairly regard a heuristic which has a better-than-chance probability of yielding a correct inference as at least minimally reliable. But on SM only those inferences and decision procedures which are sound logical or inductive rules can be counted as rational.

Theories of human rationality must also be grounded in the psychological structure of human beings. Agent psychology is fundamentally relevant for the reason that as finite creatures, our cognitive structure is crucial to the *use* of our finite computational resources. Agent psychology is ignored by all versions of SM.

The computational intractability inherent to idealised models of normative rationality follows from the inability of SM to handle these five features. Since idealised rationality (in the sense outlined by SM) fails to account for these crucial features, it must be regarded as degenerate. This provides prima facie reason to regard idealisations about the mind as suspect in general.

On the other hand, a defender of SM might be tempted to argue that SM is correct and that its ceteris paribus clauses can be filled in by Hooker's five factors. However, I maintain that the five factors cannot be accurately captured in this way. The reason comes down to the fact that these five notions are inconsistent with idealised rationality. But if SM represents a simplifying rather than degenerate idealisation this should not be the case. Perfect elliptical orbits are approximated to a greater degree when perturbing factors are removed. One can suppose there is a smooth transition towards perfect elliptical orbits as these interfering forces are removed. With respect to rationality the picture is quite different. On SM, only those inferences that are formally correct are rational, period. Once one grants that non-formal judgement can be rational, one is no longer supporting the Establishment position.

In short, the five factors cannot be integrated into SM because they are the antithesis of rationality according to SM. The forces which perturb otherwise perfectly elliptical orbits imply no conceptual contradiction. On the contrary, when the interference of the five factors is removed, what remains is a model with counter-intuitive norms. Recall that SM requires deductive closure. But why should we accept this as an ideal normative standard? In what sense can it be said to be maximally rational to draw *every* trivial inference ad nauseam?

34

Why should anyone want or be required to do that?

Interestingly, it may be the case that certain forms of SM as well as the position outlined by Cherniak in the first chapter of *Minimal Rationality* can be captured as interfering factors in ceteris paribus clauses describing rational inferences. Cherniak's earliest construal of MR, (the Weak version!), following the Establishment view, held that a Minimal Agent was specified by manifesting a subset of formal inference rules as defined by idealised rationality conditions. Recall that on this early form of MR, the minimal conditions were defined by soundness, appropriateness and feasibility. These factors, I think, could be described as perturbing forces (as an analogy to the 'orbits' example), since there is no inconsistency implied between their acceptance and SM. In both cases, rational ascription is based solely on the manifestation of sound inference rules. As his theory develops, however, the requirement for only formal rules is dropped. As I have remarked earlier, this move turns out to be a crucial one, although it is rather subtle. It is not even clear whether Cherniak himself is aware of this shift in position. Again, since non-formal judgements contradict assumptions about including only deductive and inductive inference rules, it seems implausible to suppose that all of Hooker's five factors could be incorporated into SM.

Quus-Plus Revisited

Turning back to the discussion of Kripke, it seems that there may be, after all, reasons to doubt the plausibility of some idealisations about the mind. The Argument from Scientific Idealisations (ASI) that I proposed held that making idealisations about the mind seemed no worse than making idealisations about any other domain of scientific inquiry. The trouble is that such an argument proceeds from the initial plausibility of theoretical idealisations *in*

35

principle, to the rather less plausible tacit assumption that any idealisation is prima facie justified, when in fact, with respect to cognitive systems, it is just the opposite. Again, by focusing only on the simplifying idealisations in science, we get a partial and distorted picture.

Both Cherniak's and Hooker's arguments show how the highly idealised theories of rationality are flawed. To be fair to his critics it should be noted that Kripke does not provide good reasons to reject idealisations about minds. Nevertheless, I think the reasons for such doubts are there for the taking. If the preceding analysis is correct then Kripke is defended from certain critics.

The importance of Hooker's distinction between degenerate and simplifying idealisations becomes evident when the Argument from Scientific Idealisations is elucidated. The threat posed by this argument should not be underestimated. Superficially it appears that Minimal Rationality assumes that theories of mind must be different in kind from other scientific theories. If true this would be a mysterious position indeed. The beauty of Hooker's distinction is that it is not the *theory* of mind which is distinguished from other scientific theories, but rather, a particular theoretical aspect. So the deficiency of undue idealisation was not found to be idiosyncratic of theories about the mind. Degenerate idealisations have been found in a variety of disparate scientific theories with theories of rationality being but one of many examples.

Concluding Remarks on Idealisations

Although to my knowledge he makes no specific mention of Kripke's quus-plus, Cherniak offers an example in chapter four of *Minimal Rationality* which is similar in form,

if not purpose. Suppose there is an agent who follows a deductive system different from ours, such that, one of their rules of inference cannot be used for inferences involving more than 1,000 logical constants. Call this rule *moquus tollens*. Again, like Kripke, we may ask what facts about our past use tell us whether we are following such a rule. Cherniak notes that it is likely that such an agent could perform all applicable, feasible, sound inferences by means of an unsound, or even inconsistent system. As long as the quus-like rules yield the same answers as the rules of logic, no difference in behaviour will be detectable. I certainly cannot remember the last time I made an inference involving more than 1,000 constants!

In fact, Kripke's sceptical problem may have some bearing on reality after all. Some of the psychological research indicates that the way we follow certain rules depends on the circumstances of its intended application. Cosmides has argued that common errors made during the Wason selection-task are most prominent when the task is presented to subjects as a formalised problem. When the same task is given in a social context, namely, in the context of a permission-obligation schema, performance improves markedly. Cosmides takes this to mean that we tend to solve certain problems without executing some internal logical calculus. In this case, there is presumably yet another heuristic (say the 'permission schema heuristic') which is used to relate the notion of duty and obligation in a social contract. Cherniak writes that the '1,000 constants' example is uninteresting but I think he has unknowingly raised Kripke's sceptical problem in a different way: why should we think people follow a deductive rule like modus tollens when they do the selection task? Cosmides' research suggests we do not.

1.5 Summary and Conclusions

Time to sum up. So far I have argued that Cherniak's theory of the Minimal Agent is a more realistic, useful model than its idealised rivals. The issue of vagueness has been addressed, in reply to critics as well as to expand on the initial theory. As mentioned before, though, a full explication of what 'some' amounts to will be a massive philosophical and scientific undertaking. The other major issue dealt with the controversy over theoretical idealisations. Cherniak and Hooker's criticisms of idealisations about minds were used as a reply to the 'Science as Idealisation' objection. Finally, the arguments developed against such idealisations were turned back at critics of Kripke's.

Cherniak's theory of the Minimal Agent represents an important advance over the Establishment View's highly idealised models. Cherniak's fundamental contribution is the weakening of the normative requirement for agents to act in perfect faithfulness to logic. The first important objection to his position revealed a need to expand upon the vagueness of the 'some' clause of the Minimal Normative Ideal. The explication of 'some' was partly addressed along with the reminder that the highly contextual nature of rationality supports the view that an explicit definition or rendering of the extension of 'some' is neither possible nor desirable. In the second chapter, some further remarks on this issue will be made. In particular, Cherniak's notion of tradeoffs between features of inferential systems will be expanded upon. I intend to argue that the selection history of an inferential system may be of as much importance to the project of the explication of the 'some' clause as is sociology. Cherniak's notion of epistemic tradeoffs is a useful tool, but, like his treatment of the 'at least some' clause, it lacks precision. I argue that Cherniak's notion of tradeoffs needs to be

expanded beyond the two examples (computational speed and reliability) he cites. Other features or properties of inferential systems may also be traded off due to evolutionary pressures. Hence this discussion will look at whether and to what extent an organism's selection history determines the categories and results of these epistemic tradeoffs.

The second main objection to Cherniak is, admittedly, imagined to some extent. That argument was constructed out of Pietroski and Rey's criticism of Kripke and might be thought to be an unfair characterisation of their position. Certainly my use of their remarks went well beyond their intentions. But since the rejection of theoretical idealisations about the mind is a fundamental assumption within MR it seems plausible that other critics might be tempted to challenge Cherniak on these grounds. Hence the anticipation of the imagined argument (ASI) is meant as a preemptive move against future critics of Minimal Rationality. Hooker may also have anticipated the need for further support here, and as he has shown, there are at least two distinct kinds of theoretical idealisations. The final move was to show why idealisations about the mind are identical in kind to the sorts of theoretical degeneracies found in science as the $c = \infty$ example illustrated. Finally, the discussion of idealisations was seen to have implications elsewhere in philosophy, namely, with respect to Kripke's sceptical argument about rule-following.

It seems that if what I have argued is correct, a different approach to Kripke's Wittgenstein Paradox is needed. Cosmides' research suggests informal decision-making procedures are a well entrenched inferential tendency. Perhaps Kripke's sceptical question can be finessed if we accept that "humans do not reason logically but adaptively." (Clarke 1996, 147)

Chapter Two: Truth, Consistency, and the Notion of Epistemic Virtues

Introduction

This chapter is primarily concerned with Stich's views as given in *The Fragmentation of Reason* (FR) as they relate to Cherniak's theory of Minimal Rationality (MR). In spite of a superficial similarity between their respective positions, Stich's thesis will be shown to be incompatible with Cherniak's. Stich draws three conclusions which I shall argue are at odds with Cherniak's point of view. First, Stich argues that epistemic norms must be construed relative to an epistemic agent's community; next, he rejects entirely the notion of epistemic virtues, such as truth; finally, he questions the plausibility of the idea that natural selection shaped our inferential systems. Each of these suggestions are unwelcome from the perspective of Minimal Rationality. Without some notion of epistemic virtues, it is hard to see how there could be any justifiable standards through which even minimal normative epistemic claims could be made. For Stich, reason is parasitic on culture and discovering our norms only tells us about our biases. My response will be to argue that Stich is committed to a kind of relativism that is not desirable and that his attack on the instrumental value of epistemic virtues leads to a *reductio ad absurdum*. As will be shown, these replies to Stich do no violence to Cherniak's position. On the contrary, the notion of epistemic virtues will be quite useful to Cherniak and his followers. Cherniak also makes explicit his allegiance to some kind of evolutionary argument regarding the development of our cognitive processes.[1] Further support for this assumption will be provided.

The critique of Stich I intend to pursue should not be regarded as merely a partisan defence of Cherniak's views. There are good reasons for anyone to reject the conclusions

Stich draws regardless of their attitude towards Minimal Rationality. Indeed, the differences between their positions are not at all obvious at first glance. Hence, a secondary goal will be to distance Cherniak from Stich's least appetising ideas.

2.1 Stich's Assault on Truth

In this section I shall explain and respond to Stich's arguments against the view that truth should be valued as an epistemic virtue. Through arguments raised by several critics, and a few objections of my own, I hope to show that Stich's position does not square with certain plausible intuitions about knowledge. From a successful defence of truth as an epistemic goal worth pursuing, the more general notion of *epistemic virtues* will be developed. Thus the main point of this section will be to argue for the idea that truth is but one of several desirable virtues that inferential systems might possess. The value of this notion to Minimal Rationality will become apparent during a subsequent discussion of Cherniak's ideas about trade-offs in heuristic-building and inferential system design.

To begin, in chapter five of FR Stich makes a remarkable claim: "once we have a clear view of the matter, most of us will not find any value, either intrinsic or instrumental, in having true beliefs." (Stich 1990, 101) If he is right this would be particularly devastating for reliabilist epistemic accounts of rationality and truth-linked accounts of what it is to reason well. Stich goes so far as to wonder whether knowledge, when construed as Justified True Belief , is of any value at all, for "if neither truth nor justification is valuable then the value of knowledge itself is brought into question." (Stich 1990, 101) Stich mentions that his position is attractive because it provides a good reply to the epistemic sceptic who denies

the reality of all knowledge. If knowledge, or (only) true belief has no value then the epistemic sceptic's objection has no teeth. If we do not care about having knowledge we will not care whether or not we actually have any knowledge. But a bit of back peddling is in order before we join Stich and abandon any interest in knowledge. Firstly Stich has to establish that there is no sense in valuing true belief.

Stich constructs arguments against the idea that there is any intrinsic or instrumental value in having true beliefs. But in order to understand Stich's arguments it will be necessary to explain his notion of an 'interpretation function'. For Stich, an interpretation function works like this: an interpretation function is what "pairs psychological states with propositions, or truth conditions..." (Stich 1990, 105) Stich defines alternative causal/functional/historical relations of brain states to the world as TRUTH*, TRUTH**, and so on. The inferiority of truth over some alternative word-world links turns on the banal fact that having true beliefs does not always maximise desire satisfaction. Surely, at some time or other, every believer has failed to achieve a degree of desire fulfilment that could have been attained had the agent's beliefs been different in some way. To illustrate Stich gives the example of Harry whose true belief leads to his own demise. Harry correctly believes his flight leaves at 7:45 a.m.. He arrives at the airport in time and catches his flight, however, his plane crashes and he dies. Had he only (falsely) believed that his flight left later-say at 8:45 a.m.-he would have missed his plane and survived. Presumably Harry was not suicidal, hence, his true belief led to a frustration of his desires. Therefore, for a given true (and commonsense) interpretation function, there is some other function which maps the agent's brain states onto whatever sets of beliefs would have *optimally* fulfilled that agent's desires.

42

For Stich, "the only obvious complaint to lodge against many of these alternative schemes for nailing words on to the world is that they do not happen to be the scheme sanctioned by our commonsense intuitions." (Stich 1990, 115) Poor Harry.

Now, if by 'intrinsic value' it is meant 'always beneficial' then as an argument against the intrinsic value of truth, the persuasion of Stich's example must be granted. But even if this pragmatic interpretation of value is granted it nevertheless seems that his example does not bear much relevance towards instrumental accounts of the value of truth. This outcome is somewhat puzzling given that Stich writes (1990, 122) that his example is meant to dispatch with the *instrumental* value of belief. To show that truth has no instrumental value, Stich would need to demonstrate that having true beliefs was not better than having false ones in general, and he has certainly not done that. But I am getting ahead of myself here.

Before proceeding further with Stich's arguments, it should be mentioned that I shall focus only on his remarks concerning the instrumental value of truth. I want to eschew discussion of intrinsic value for two reasons. First, I suspect that any assignment of intrinsic value to a thing can be explained in terms of its instrumental value. For instance, sometimes one hears things like "the environment is intrinsically valuable above and beyond any use to us." But this seems like a strange thing to say. Is not the environment valuable to us because it sustains our lives, is beautiful, is worthy of scientific study, and so on? If the environment did not provide any of these things-sustenance, beauty, and so forth-I doubt that anyone would find it valuable in any sense. If this is right then the notion of intrinsic value is left without any work to do. Secondly, and more importantly, with respect to the evolution of

inferential systems, only features which are of instrumental value should be of any interest. There is no sense in talking about the intrinsic value of an adaptation. So, since I find the idea of intrinsic value to be troublesome, I have little interest in defending it here. That being said, Stich makes three major points regarding the instrumental value of true belief. (Stich 1990, 124)

2.2 Three Points and Three Replies

First, in order to establish the instrumental value of truth, Stich would say that it is not enough to show that true beliefs are superior to false ones. What must be shown is that true belief is superior to the "indefinitely many alternative categories of belief (TRUE*, TRUE**, and the rest), which may by and large be counter-intuitive." (Stich 1990, 124) The proper response to this first point is that it is not correct that truth must be shown to be superior to the *indefinitely* many alternative categories. As will be shown, many of these alternate categories are not viable options. Cherniak's point about not being obliged to do what we cannot do cuts off any alternatives that are not already built into or teachable to our cognitive systems. If this is so, the burden is then on Stich to determine how close truth is to the most useful and viable of design options. I would agree with those who suggest that in fact it might be very close indeed. Truth should be dear to the hearts of any organisms that want to have their desires satisfied. I cannot eat the berry if I do not believe there *is* a berry. *Truth****, Gruth, and Truth!*

Critics of Stich, including Bach, Jacobson, and Lycan have also raised serious doubts about the ability of alternative interpretation functions to do what Stich claims they can. In

44

what follows, I shall look at some of these objections with the intent of showing that only truth, or a function which is very truth-like, could be a serious candidate.

Lycan has coined the term 'GRUTH' to stand for an interpretation function in which there is some belief p such that although the agent believes p it would be more advantageous for that agent to believe not-p. Notice now that creatures that have grue (in Lycan's sense-not Goodman's!) beliefs are more adaptive than those that do not, "and by hypothesis, some grue beliefs are not true. Thus if evolutionary advantage is what matters, we should want our beliefs to be grue, not true." (Lycan 1990, 202) Yet some fine tuning of Lycan's notion is in order. As Bach (1992) correctly points out, the trouble with Lycan's notion of GRUTH is that it applies only to a *specific* belief. For Harry to have beliefs which are GRUE only one of his beliefs needs to satisfy the requirements of GRUTH. Hence GRUTH represents only a modest improvement on truth-an improvement, that is, of only one belief. Bach's suggestion is to generalise Lycan's notion of GRUTH in order to get at the alternative to truth which is the best of all possible alternatives. That notion, TRUTH!, in contrast to GRUTH, just is the optimal belief set for anyone. TRUTH! then, and not GRUTH, would have the maximal adaptive advantage.

Obviously, though, there would be many problems in making TRUTH! a normative ideal since it is likely that each agent would have very different sets of beliefs under TRUTH!. For instance, in a given situation it may be TRUE! for me to believe that p and also TRUE! for you to believe not-p. To see why just imagine the following scenario: Suppose I am at a demonstration and I believe I see someone protesting against cuts to Unemployment Insurance. It is a good thing that I believe this since this protestor is my ride

45

home. Now suppose that Cretien also sees the protestor, approaches him and for some reason, chokes him. The ensuing wave of bad publicity ruins Cretien's credibility as a "nice-guy" politician and dashes his hopes for re-election. If only the Prime Minister always had beliefs which were TRUE! this unfortunate situation would never have happened. Here, Cretien's relevant TRUE! belief is "There is no protestor." If his beliefs were TRUE!, he would not have choked the protestor because in his mind the protestor would not have existed. It is easy to see that there may be a great number of similar cases in which it would be advantageous for different people to have drastically different beliefs *about the same thing*.

Another, larger, difficulty for TRUTH! arises from certain considerations of feasibility which restrict us from pursuing beliefs which are TRUE! (or for Lycan GRUE). Lycan notes that "gruth is not a viable design option. Neither natural selection nor a beneficent and skilful Mother Nature could have fashioned a cognizer to have consistently grue beliefs, since the vagaries and vicissitudes to people's life histories vary unpredictably and far too widely across the species." (Lycan 1990,204) (Note that one should substitute 'GRUTH' for Bach's corrected notion of 'TRUTH!' here.) Exactly. The difficulties Mother Nature must face when involved in building a TRUE!-belief generating system must be at least as great as those involved in building any optimally designed organism, (just imagine one, say a nuclear-powered butterfly equipped with a cloaking device). Since the Cretien example above shows that there could be no *species* specific TRUE! belief set, that seems to rule out natural selection as a possible mechanism. Consider also how the outcome of innumerable future events would determine which beliefs are TRUE!, as with, say, knowing

46

when to buy or sell stocks. It seems that someone with a maximally TRUE! belief set would have to be clairvoyant! TRUTH! then cannot be taken as a serious normative rival to our ordinary notion of truth. But what about some weaker version of TRUTH!?

Lycan grants that it may well be that selection produces approximations of grue-belief generating systems but that the most grue-like approximation is probably very close to truth itself. Lycan, Stich, and possibly Bach would agree to regarding the notions of truth, GRUTH, and TRUTH! as connected on a continuum of possible truth-like notions. In that case the issue becomes determining how close our actual belief forming mechanisms approximate gruth. All would agree this is a "brutely empirical question" (Lycan 1991, 205) lacking any satisfactory answer at this time. Yet, even if all Stich has offered is the conceptual possibility of an alternative to truth, and there is no accompanying detailed proposal, it becomes very hard to take his suggestion seriously.

The Rejection of Selection

The second main point Stich makes is his rejection of arguments about the evolution of true-belief producing systems. For Stich, claims to the effect that our beliefs are more often true than not or that our cognitive processes are, in general, reliable are panglossian. Since natural selection is not the only force of evolutionary change, the effects of random sources of change cannot be ruled out.

Genetic drift, for example, occurs when a gene becomes fixed in a population while its competitors disappear. Drift can lead to the elimination of a more fit gene and the fixation of a less fit one. Also, the influence of pleiotropy, that is, when one gene affects more than one trait, can also result in a less than optimal design. Some genes may have positive effects

towards one system, but negative effects towards others. In such a case, it may not be possible to possess a particular advantageous trait without giving up some other advantageous trait elsewhere. Stich cites the example of arctic animals whose white camouflage results from the same genes which also cause eye problems (albinism).

Furthermore, even if it is true that selection plays a dominant role in the shaping of our inferential systems, Stich argues there is no reason to think that such systems must necessarily approximate optimal design. There are many reasons why selection may not produce optimal or near optimal designs. For example, the required mutations may never happen to occur, since the possibilities open to selection are restricted by the basic materials available. Having built in wheels or rocket engines might approximate an optimal design for a speedy predator, but that does not mean that there will be four wheeled cheetahs one day. Gould illustrates a similar point nicely with an example about orchids:

> Orchids have evolved an astonishing variety of "contrivances" to attract insects, guarantee that sticky pollen adheres to their visitor, and ensure that the attached pollen comes in contact with female parts of the next orchid visited by the insect...The message is paradoxical but profound. Orchids manufacture their intricate devices from the common components of ordinary flowers, parts usually fitted for very different functions. Orchids were not made by an ideal engineer; they are jury-rigged from a limited set of available components.[2] (Gould 1980, 20)

Random drift is another possible evolutionary factor to consider. Random effects can eliminate otherwise favourable mutations thereby overriding the "force" of selection. Just because a change in gene frequency results in a fitter organism does not mean that the mutation will necessarily become dominant within a given species. Some natural calamity- say a meteor strike-could wipe out an otherwise fitter subpopulation before the new adaptation spreads to the rest of the species.

48

Taken together these points show that selection cannot a priori show that human inferential tendencies, including the tendency to value truth, were selected for because of their close approximation to optimal design. Yet it seems that all this provides for Stich is a prima facie argument against the idea that the evolution of reliable inferential systems was primarily the result of natural selection. In fact, all Stich has shown is that it is possible *in principle* that certain traits or gene frequencies responsible for cognition were fixed for reasons other than selection. What Stich does not provide are *specific* arguments aimed at showing why we should think that our brains evolved as a result of Genetic Drift or Pleiotropy. For a lack of a substantive argument on either side, Stich chooses to remain agnostic. (Stich 1990, 124)[3]

One way of providing a rejoinder to Stich would be to give an argument as to why we should think that human cognition is most likely the result of selection and not some other process. It may be the case that he has not looked hard enough. I think there are at least prima facie arguments to support that view.

Reply to Stich

J.S. Mill pointed out that brain size-simpliciter-is not a reliable indicator of intelligence, otherwise one would expect whales to be Cambridge dons. Even *within* a species comparisons of brain size are a crude parody of real measurement. In humans there is a wide range of normal brain weights and comparison is made more difficult because brain size is related to body size; a relatively smaller brain usually means only a relatively smaller body. Across species the matter is further complicated by the fact that larger animals tend to have relatively smaller brains, hence a comparison based solely on the ratio between brain

weight and body weight will not work either. Nevertheless, the size of a brain, corrected for size differences both within and outside of a species may reveal much about cognitive abilities. Gould (1977) has previously pointed out the right way to interpret differences in brain size: "The correct criterion is...the difference between actual brain size and expected brain size. To judge the size of our brain, we must compare it with the expected brain size for an average mammal of our body weight." (Gould 1977, 183) This comparison results in a value known as the *encephalization quotient* of a species or 'EQ'. EQ measures brain weight as compared to an "expected" value of an average mammal, so for instance the average expected EQ for a modern mammal would be 1.0. Mammals with an EQ higher than 1.0 would be expected to possess greater cognitive abilities. With this notion in mind it can be shown that there is some evidence which supports the idea that selection played a crucial role in the increase of brain size if the assumption that there is a connection between brain size (corrected for the various factors mentioned) and intelligence is allowed.

Gould argues that evolutionary feedback-loops result when there is intense selection pressure between predator and prey. To illustrate, Gould (1977, 189) cites research by Jerrison which compared the encephalization quotients of Carnivores in the Tertiary Period (from 70 million years ago to the present) and the Herbivores they most likely preyed upon. When encephalization quotients are compared in this manner the phenomenon of evolutionary feedback presents itself:

Encephalization Quotients:

	Herbivores	Carnivores
Early Tertiary (archaic)	0.18	0.44
Early Tertiary (advanced)	0.38	0.61

50

| Middle to late Tertiary | 0.63 | 0.76 |
| Modern | 0.95 | 1.10 |

What these numbers indicate is that brain weights of herbivores and carnivores increased in concert. It is assumed, and plausibly so, that having a high EQ is related to the ability of predators to catch their prey and for the ability of their quarry to escape. Presumably it is more difficult to hunt successfully than to escape capture, so predators tend to have larger brains (in the sense outlined by EQ). Gould's explanation for the dual increase over the Tertiary period is that either the predator or the prey increased in brain size as a result of some favourable mutation. Henceforth there was an intense pressure on the other side to also increase in their cognitive capacities if they were to survive. What resulted is akin to an arms race in which each side grows stronger and stronger although a balance of power is maintained.

What makes this example philosophically interesting is that it is unlikely that the evolutionary feedback-loop described above is the result of Pleiotropy or Drift. The increase in brain weight is the result of the relationship between different species but drift and pleiotropy affect only gene frequencies within a species. Gould provides additional evidence to support this view. Until the rise of the Isthmus of Panama, South American Herbivores were isolated from more advanced Carnivores. When North American Carnivores finally invaded that continent the native Herbivores were quickly wiped out. Throughout the pre-invasion period Herbivore encephalization quotients remained constant, presumably because predatory roles were filled by marsupial carnivores with low encephalization quotients. Hence in reply to Stich, it seems that there is evidence to support the notion that the evolution of mammalian brains was strongly influenced by selective pressures.[4] If the

NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was microfilmed as received.

52

This reproduction is the best copy available.

UMI

be harmful. Stich (1990, 118) takes the standard interpretation of what intrinsic and instrumental value are. Things are valued intrinsically if they are valued for themselves. Things are valued instrumentally if they are useful in leading one towards some other valued thing. Yet just because truth is not always useful it hardly follows that truth is not instrumentally useful. Having a Straight Flush in Poker usually helps facilitate winning the round. However a Straight Flush will fail you if one of your opponents happens to be lucky enough to have a Royal Flush. In fact, in such a case a good hand like a Straight Flush would be downright harmful since your betting would likely be much more unrestrained given the supposed strength of your hand. Nevertheless I think I am justified in thinking that Straight Flushes are instrumentally valuable when I play Poker.

It may be that Stich-and my arguments above up until now-have assumed that virtues must be reliable. But Stich's requirement that virtues must always represent optimal strategies seem excessively idealised. Defenders of epistemic virtues, such as truth, should not feel compelled to show that truth is a reliable virtue. To see why, consider Millikan's (1984) criticism of reliabilist theories of representation. Millikan points out that it is selective advantage, and not reliability, which is more relevant to how representational systems are constructed. Adaptive functions may be unreliable in two ways: they may be instantiated only rarely (few sperms ever fertilize an egg) or only under rare circumstances (the vomit reflex). It would be a mistake to think that the design of sperm is somehow deficient just because only a tiny fraction ever fertilize an egg. The camouflage of the Woolly Caterpillar may provide only a slight competitive edge over its uncamouflaged cousins, but even a slight advantage could make the difference between evading a predator

and becoming someone's lunch.

A Moving Target

Confusion arises from the conflicting claims Stich makes regarding his own goals in Chapter 5. As I have mentioned before, he begins by claiming that "once we have a clear view of the matter, most of us will not find any value, either intrinsic or instrumental, in having true beliefs. (1990, 101) Elsewhere, though, Stich cautiously backs away from this claim as when he writes "...nothing I have said comes even close to a knockdown argument against according intrinsic value to having true beliefs." (1990, 120) Instead Stich takes the view that those who value truth are epistemically conservative, "letting tradition determine their cognitive values without any critical evaluation." (1990, 120) He similarly hedges his remarks about instrumental value as when he admits that "since many people probably value many things...instrumentally, I'll make no attempt to argue that having true beliefs *could not be* instrumentally valuable" (his emphasis). Perhaps he was trying to rouse interest in the reader by initially exaggerating his claims. Or maybe he realised only later that his arguments would not support his thesis as originally stated. Whatever the reason, Stich himself seems to be acknowledging that the most extreme interpretation of his claims cannot be right.

What Stich does say is that if Harry had falsely believed that his flight left at 8:45 a.m., he would have lived. Stich dresses this assumption up to resemble an alternative interpretation function, namely, TRUTH****. (1990, 123) But Stich's alternate mapping scheme TRUTH**** just is the commonsense one plus the false belief that the plane is leaving at 8:45 a.m.; in other words Lycan's GRUTH. More formally, Stich's argument

54

against the instrumental value of truth seems to boil down to:

P1 Knowing the truth is sometimes harmful
P2 There is some alternative to truth, TRUTH*****, that is less harmful than truth.
P3 An epistemic virtue is the least harmful of all alternatives.
C1 Truth is not an epistemic virtue.

Presumably since an alternative to truth that is less harmful has a greater instrumental value, having true beliefs cannot be the best epistemic strategy. But consider now the following:

P4 Eating healthy foods and exercising regularly is sometimes harmful: I could be hit by a bus on the way to the gym.
P5 There is some alternative to health, HEALTH*****, that is less harmful than health.
P6 A virtue must be the least harmful of all alternatives .
C2 Health is not a virtue.

There are two points to make here, although the first has already been made but could use some emphasis. The first point is that it is not enough to simply imagine there are better ways to form good beliefs, attain health, happiness, kindness, and such. As argued before, the burden is on Stich to *specify* exactly what at least one of these alternatives might be like. Clearly he has not done this and as I have argued above there are even good reasons to think that-with respect to truth anyway-there cannot be any other viable alternatives. The second and more important point to make is that if Stich's arguments about truth were taken seriously we should abandon the very *concept* of virtue! Since his argument as presented in the above schema could be applied to virtually anything anybody values, instrumental or otherwise, this observation might be thought to constitute a reductio ad absurdum argument against Stich. Jacobson (1992) makes similar remarks concerning the value of truth[6]:

> Presumably, we could cook up a large number of alternatives-Health*, Health**, and so on-and contend that our acceptance of health as an intrinsic value is not based on a consideration of these alternatives. If the acceptance of truth as an intrinsic value is objectionable for the reasons Stich gives, so is the acceptance of health, and presumably anything else we might accept as an intrinsic value. (Jacobson 1992, 340)

Note that for my purposes 'instrumental' should be substituted for 'intrinsic'.

Stich is right about one thing, namely, that there should be an argument as to why truth has an instrumental value. The form it should take is something like: ceteris paribus it is better to have true beliefs in the long run. Better than what? Better than all of the feasible alternatives? How about "better than all of the feasible apparently appropriate alternatives?" Having a good gambling strategy will not guarantee success, but, it will make you win more than you lose. A better-than-chance gambling strategy which wins 51% of the time will fare worse when pitted against a better strategy that wins 52% of the time. But just because there are other strategies that yield returns greater than 51% does not mean that the 51% strategy is not instrumentally useful. So an instrumentally valuable strategy need not be an optimal strategy, nor even the most optimally-feasible-apparently appropriate-strategy. Whether or not a heuristic, say, is instrumentally valuable depends on what desires you are interested in fulfilling. If you wish to maximize your return while gambling, the 51% strategy is not instrumentally valuable. However, it is of great instrumental value if your goals are more modest and you only wish to turn a profit.

Thus the main point: the notion of epistemic virtue, at least with respect to truth, is legitimate. A virtue is just something we would like our inferential systems to approximate or something we would encourage other people to acquire. The entire discussion was meant to give one such core example, truth, a solid backing. But no one need say that truth is the only, or even the primary epistemic virtue we should desire. Hence, truth need not be optimally realised, but rather, truth should be valued to the extent towards which it is useful. In the next section I shall consider the idea that there is a multiplicity of inferential virtues.

56

2.3 Epistemic Virtues and Evolutionary Tradeoffs

So far I have tried to provide replies to Stich's arguments against the epistemic virtue of truth as well as demonstrate how his views could lead to the (absurd) denial of any virtue whatsoever. This is thought to be needed because the notion of epistemic virtues can be related to Cherniak's notion of 'epistemic trade-offs'. Epistemic tradeoffs are the result of a delicate balancing between various competing features, properties, or aspects of inferential systems. Cherniak only explicitly mentions truth preserving outputs and computational speed, but with a little imagination, it is easy to think of others. If we suppose that the evolution of inferential systems is roughly analogous to the design of any complex functional system, say a bicycle, then an epistemic virtue like truth is somewhat like the feature of top speed as it figures into the design of a bicycle. Different designs of bicycles may more or less favour different virtues or features to the extent that certain tradeoffs, say between speed and weight, have to be made. Hence the degree towards which any virtue-in the sense of a feature of some functional system-is instantiated, need not be maximal; not all designs of bicycles are as fast as others even though one of the features a bicycle typically has is that it allows the rider to travel more quickly than usual. In short, virtues need not always be the optimal realisation of all possible alternatives. One useful analogy to keep in mind here is that of a good gambling strategy. Smart betting is no guarantee for success, but, supposing fair play, it's the best thing going. I think that the right interpretation of Cherniak supports this view.

Complexity theory provides a distinction between computational problems which are

practically feasible and those which are only theoretically so. Formally correct deductive processes may be so slow in some cases as to yield computational paralysis. Therefore, messy heuristic inferential strategies are not necessarily less desirable than those which are formally correct.

The core of Cherniak's point is that *reliability*, in the sense of generating true outputs, is not the only, or paramount, virtue to be had by an inferential system. Other virtues such as computational speed may be just as important. Yet if we accept only these two desideratum an ideal inferential system will be one with the best tradeoff between speed and reliability. It should not be thought that there is one and only one ideal tradeoff to be made that would apply to every situation or individual. In humans, the kind of tradeoff made between speed and veracity would fluctuate depending on what inferences are being made under what contexts. So, for instance, a tradeoff that favoured speed would be more appropriate while making inferences on the run, while a slower more deliberative approach would be preferable as long as time allowed. So such tradeoffs need not and probably are not restricted to the confines of our inherited cognitive traits. Thus a certain amount of built in flexibility is desirable in a cognitive system. Certain cultural and individual idiosyncrasies may further shape the exact character of what kinds of tradeoffs a particular person or persons possess. Further, it should not be thought that computational speed and reliability are the only two epistemic virtues to be evaluated.

The points made above explain why our inferences cannot be deductively perfect. Note that making speed a virtue on par with reliability is not ad hoc. According to evolutionary theory we should not be surprised at all to learn that computational speed is an

epistemic virtue as well. If our inherited and learned cognitive strategies do converge towards roles which help facilitate our survival, then truth preservation cannot be regarded as primary. It should not be assumed that unrelenting pursuit of truth will always produce the best results. As strange as it may sound, the truth must give way to other epistemic and non-epistemic virtues if it would help us attain our goals. This is not to say, as it might be thought, that I now agree with Stich that truth should be replaced by some other notion. Unlike Stich, I hold that it would be best for our beliefs to optimally realise truth. The trouble is that this becomes increasingly difficult to accomplish given other pressures; maximal approximation of truth is traded for speed, and so forth. Stich, in contrast, would prefer to question the value of truth altogether.

One way of thinking about this interpretation of Cherniak is represented as Figure 2 below:

Fig. 2: Cherniak's Notion of Tradeoffs
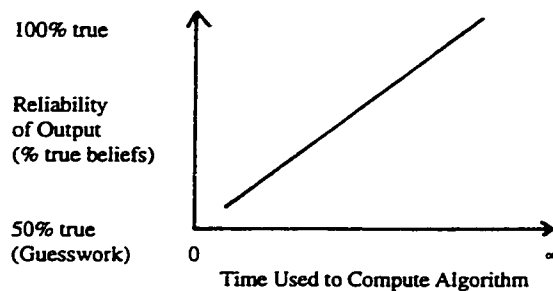
A few words of caution should be inserted here regarding the use of such graphical representations. The graph is highly abstract and is meant to be taken more as a guide for the imagination than as a quantitative representation. For example, there is no way of determining what the slope of the "inference" curve in fig 2 is, or, exactly how values for X and Y are to be quantified in the first place; and it would be incredibly naive to think that

they could be here.

But even with this warning firmly in place such figures may still be helpful. The idea is that tradeoffs between a multiplicity of epistemic virtues results in rough hewn boundaries of a *reason-space*. This space represents all the possible combinations of the various epistemic virtues that a Minimal Agent's cognitive system could occupy. Considerations of feasibility, appropriateness, consistency, and truth would be built into the boundaries of reasonspace for the Minimal Normative Ideal. For another helpful pictorial analogy, imagine that the reason-space occupied by the Minimal Descriptive Limit. It would be necessarily somewhat smaller than the Minimal Normative Ideal. Actual agents would be expected to place somewhere within the space defined by the upper and lower limits of the Minimal Descriptive Limit.

Allow me now to make a few comments about the plausibility of adding other concepts besides truth and speed to the list of epistemic virtues to be considered when thinking about reason-space. Although computational speed and truth are worthy epistemic tools, surely there are other inferential virtues that should similarly be accounted for. Several possible candidates come readily to mind, namely, energy expenditure, or, the construction cost of an inferential system. But there could be many others including memory size, memory retrieval, and consistency. Goldman (1986) comes close to this sort of view in his discussion of the power, speed and reliability of cognitive systems. Goldman suggests that power could be thought of in several ways including the raw number of beliefs a system can generate or perhaps the percentage of problems a system can correctly solve. Like Cherniak, Goldman fashions the concept of reliability in terms of the percentage of generated beliefs

60

that are true. Since belief acquisition can be regarded as a kind of inference making (from sense data and other beliefs) it might be best to not think of the virtue of power in terms of the percentage of problems successfully solved. Otherwise power and reliability end up meaning the same thing. Nevertheless, Goldman's other suggestions should be added to the growing list of cognitive virtues.

To return to the bicycle analogy, one might wonder what features are desirable in a bicycle and to what degree should they be instantiated? The answer to that question depends on what specific function the bicycle is intended for. So mountain bikes (say) need not be as fast as racing bikes. Yet although some bikes will share more of a family resemblance than others, some things will apply to any bicycle design. Since it has to be rideable, considerations of the dimensions of the frame, weight, and the mechanics of the gears all immediately come into play. Also on this list is speed, serviceability (reliability), handling, cost, and perhaps even an aesthetic quality. Similarly it can be supposed that inferential systems will display similar compromises between competing features as with say reliability, power, construction cost, speed, active memory size and maintenance of consistency. Notice that here the virtue of speed (S) is distinguished from the requirement of computational time (CT). CT is the time required to complete a given inference (given an average cognizer), while S represents how quickly an inferential system can come up with an answer. CT is analogous to the size of the race track while S is analogous to the speed of the runner. Similar features must be considered when designing a bicycle with the overall function of the resulting design as the guiding idea.

Notice, however, that unlike cognitive systems, bicycles are human artifacts and not

61

biologically natural systems. Evolved cognitive systems may never approximate optimal design. Unlike human designers and engineers, Mother Nature is in Dawkin's phrase a "Blind-Watchmaker". Nevertheless with respect to the evolution of inferential systems, the metaphorical use of teleological terms and notions expresses the same underlying phenomena: giving-a-little-here-to-get-a-little-there. Consider cost, in terms of complexity and bulk. An otherwise fitter highly sophisticated inferential system may be at a selective disadvantage when compared to cheaper "inferior" designs which can replicate with greater fecundity. So the candidate which yields the best ratio between the speed and reliability of inferential systems on the two-axis graph could well be far more costly than some less reliable slower system. Consider, too, the notion of serviceability: more complex systems might tend to be more prone to irreparable damage from injury, illness, or construction errors. The interconnected nature of the process of generating multiple tradeoffs makes for extremely complicated interactions.

Yet another way of contrasting these considerations with the Establishment view and Cherniak would be to describe the former as 'Mono-virtuists', and Cherniak as a 'Poly-virtuist'. The standard view uses only one feature of inferential systems as the yardstick of prescriptive epistemology: truth. Cherniak's arguments from complexity theory correctly opens the path for poly-virtuists by advocating a two-axis analysis but goes no further. To this I add, why not have more than two inferential virtues accounted for? Again, against Stich, the recognition of other virtues in no way detracts from the notion that truth is valuable.

As a final thought consider a few more possible tradeoffs: reliability (R) versus

computational time (CT); power (P) versus speed (S); speed (S)versus cost ($); Long term

memory size (LMS) versus consistency (C) and power (P) versus cost ($). Note that what

is most difficult to represent is the final result of the many smaller tradeoffs (figure 3).

Fig. 3: Some Possible Epistemic Tradeoffs
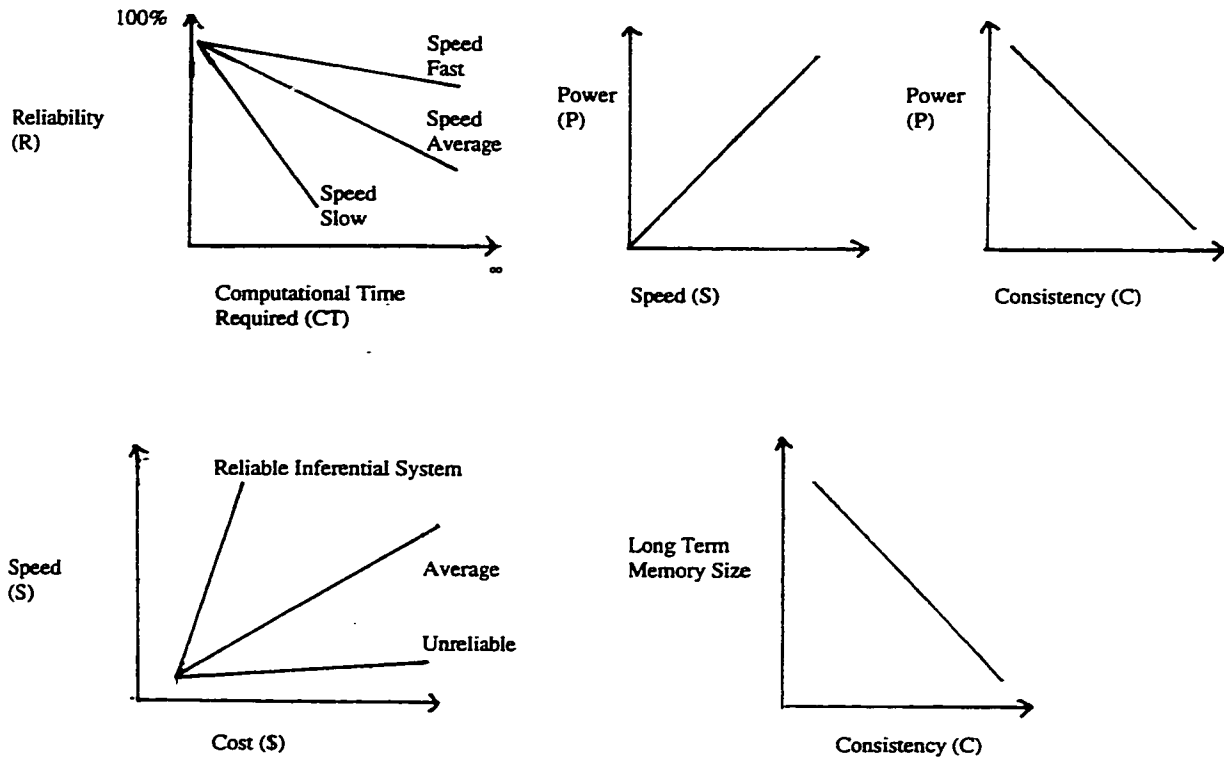
In the final section I shall zoom in on one particular inferential virtue, that is,

consistency, in part because I disagree with Cherniak's treatment of consistency in *Minimal

Rationality*. Cherniak greatly downplays the need for global consistency in one's set of

beliefs. In contrast, I suggest that consistency should be regarded as a feature, like truth,

speed, and material cost, which is also traded-off in some complex manner against other

competing features. True beliefs help us stumble through the world, but so too do consistent

ones. What good is my true belief that the berry is poisonous if I also believe the berry is not

poisonous? I shall argue that minimal consistency may also be needed to avoid a behavioural paralysis.

2.4 Consistency of Belief

The final issue to be dealt with in this chapter concerns the role of consistency in Minimal Rationality. I wish to argue that consistency bears on the preceding discussion of epistemic tradeoffs in that consistency may be fairly regarded as an epistemic virtue. Nevertheless, it is not required that agents maintain perfect consistency. Cherniak thinks that inconsistency may not necessarily be an indication of irrationality. Instead, "inconsistency may sometimes be downright healthy." (Cherniak 1986a, 99) Since Cherniak thinks that certain informal decision making procedures may be preferable to slower but more formally correct methods, an inconsistency which arises from such heuristics need not be bad at all. This observation is correct and in keeping with the notion of tradeoffs I am trying to develop. But it should not be assumed that consistency is parasitic solely on tradeoffs between speed and accuracy.[7] Another idealisation assumed by Standard Model theorists is that an agent's beliefs need to be globally consistent. Cherniak wants to show that the requirement for consistency in an agent's set of beliefs is not necessarily desirable. In short, I agree with Cherniak that the value of consistency should be judged on a case by case basis, but disagree with his treatment of the particular case he focuses on.

The Structure of Human Memory

Given that one need not adhere to logic universally in order to meet the minimal requirements of agency, one might wonder if the model allows for the possibility of agents

64

who hold many contradictory beliefs. Cherniak wishes to draw distinctions between different kinds of inconsistencies in part based on the structure of human memory. Presently, the dominant theory of human memory is the duplex theory. Although this theory is not without its critics it still represents a serviceable model. On the duplex theory, human memory is divided into two parts, namely, short term memory (STM) and long term memory (LTM). Short term memory is characterised by having a small size-usually about seven items or chunks of items, and information held in STM decays quickly although cognitive operations can be performed on information stored in STM. LTM, in contrast is thought to have a (relatively) infinite storage capacity. Here information decays slowly although information in LTM is dormant. Information held in LTM must be accessed and copied into STM before it can be used. If one thinks of the human mind as a sort of office, then STM can be regarded as the work space holding current projects while LTM is analogous to the stored files. Hence a major hurdle to global belief consistency is the necessity of accessing beliefs which may contribute to a contradiction.

Since the limitations of our short-term memory disallows evaluating the consistency of all of our beliefs at once, it is not too surprising that certain inconsistencies arise. If we could juggle all of our beliefs in our working memory-a necessity if a global consistency check is to be performed-then there would be no distinction between long and short term memory anyway. Hence the structure of human memory rules out the practical possibility of global consistency. Cherniak points out that we should not regard all inconsistencies as equally bad. Inconsistencies between very small sets of beliefs, say if the only beliefs are a and not-a, are prima facie unacceptable. However, inconsistencies in very large sets would

be much harder to detect. In such cases inconsistency may be preferable to the costs of an extensive memory search.

Memory is structured in a nested format. The way we access details about a particular subject is often like following branches on a tree. So, for instance, it would be very hard to try to remember details about the sorts of procedures I need to follow when preparing to fly an aeroplane while I also try to remember how to brew beer, and, indeed, when one does start thinking about some subject, related information usually comes to mind. For example, when I think about *Moby Dick*, the recall of details about the story tend to be interconnected. It may be that I remember that Queequeg's coffin was used by Ishmal as a life-preserver only after I have spent some time remembering other details: the name of the ship, the *Pequod*, or whatever. The explanation of this anecdotal example is well supported by current psychological data. For Cherniak, the crucial difference between long and short term memory is that "the only items available for rational processing are those currently in short-term memory." (Hooker 1994, 191)[8] The chunking structure of memories is a likely source of nested inconsistencies. Of course, once an inconsistency is detected it should usually be resolved in some satisfactory manner. Yet not every inconsistency should be automatically assumed to be worth getting rid of. It would depend on whether the costs involved in rooting out the inconsistency would be worth the effort.

Paradoxes of Consistency

Should we have consistent sets of beliefs? Are some inconsistencies allowable; are some inevitable? These questions will be front and centre in the next section. Consider one version of the Preface Paradox: Suppose I have a belief *p*: "At least one of my beliefs is

false." Is it reasonable to believe p? Well, suppose Alice believes p, then Alice seems to have contradicted herself. Consider the case where we suppose Alice's belief set contains only one other belief q. Alice believes q but her acceptance of p implies that she also believes that q is false! Acceptance of p ensures that one cannot have a consistent belief set. The problem is that it seems that it is very reasonable for any agent to accept p. Given the vast numbers of beliefs each of us has, and the fact that no one is omniscient, it is doubtful that anyone is justified in thinking that all of their beliefs are true. Hence the paradox: it seems reasonable to hold a belief which guarantees that one will have an inconsistent set of beliefs.

Cherniak has this to say regarding the Preface Paradox:

> [M]uch of the perceived paradoxicality of the Preface Paradox seems to arise if one presupposes an idealization of the agent's psychology that is very like Quine's; and, to that extent, that apparent paradoxicality can be dissipated by adopting a more adequate model of human memory. (Cherniak, 1986a, 52)

Cherniak argues that the apparent paradox follows from the assumption that agents must maintain consistency. Quine's Web of Belief metaphor, for instance, is usually thought to require an internal consistency from each strand of the web; Quine's theory makes coherence the paramount virtue to uphold. Cherniak's solution to the Preface Paradox is to deny that a lack of consistency is something we should necessarily worry about.

But I think Cherniak is too quick at accepting the consequences of the Paradox. Like Stich with respect to his arguments about evolution, Cherniak has established only prima facie plausibility for the notion that inconsistency is not always a bad thing. Whether this applies to the specific case of the Preface Paradox is another matter altogether. The trouble I have with Cherniak's explanation is that there seems to be at least one other way of looking

at this and similar paradoxes about belief consistency that preserves our intuitive notion that consistency should be maintained and which also squares with the requirements of Minimal Rationality.

Maybe there is a way to resolve any belief-consistency paradox without giving up the idea that consistency should be maintained. For reasons I shall explain, the solution is most obvious when applied to the Lottery Paradox. Consider, then, the Lottery Paradox: Suppose that I hold a ticket in a lottery in which there are 99 other tickets. Now suppose that I have the following beliefs:

1. Ticket 1 will lose.
2. Ticket 2 will lose.
   .
   .
   .
100. Ticket 100 will lose.

But given this I should conclude that *no* ticket will win. Since someone has to win, something must have gone wrong with the reasoning given above. When things are put this way, Cherniak's advice to simply ignore the problem seems itself problematic. How could he accept that an agent can believe that no one will win the lottery, for why would anyone even buy a ticket in a lottery that no one can win? If the structure of the Preface and Lottery Paradoxes are similar enough then this observation suggests that there is something wrong after all with believing that it is fine to assert "I believe that one of my beliefs is false." Nevertheless, I do think that it is reasonable to think that. It is just that one cannot justify that claim given Cherniak's account of the problem.

The solution to the Lottery Paradox rests on rethinking what the concept of belief is all about in the first place. I would argue that there is a tacit assumption about our notion of

belief which represents yet another degenerate idealisation about mental phenomena. That assumption is this: A belief is the sort of thing we must give wholehearted assent to. In contrast to this assumption it might be thought that it is more reasonable to think there is a probabilistic character to our assent for many of our even deeply held views. Goldman's remarks (Goldman 1986, 83) on this matter help greatly. Goldman, following a long tradition, calls this first notion of belief 'binary'. On the binary view, beliefs are things that a person either believes or does not. Just as a light switch can be either on or off belief is regarded as an all-or-nothing proposition. He notes that others, including himself, take the view that belief is a graded or quantitative notion. Here, assent to a proposition can be expressed with degrees of confidence: propositions range from 0 (none) to 1 (full) assent. There are many propositions that we would be inclined to believe, but with some hesitancy. Cases in which people say things like "I don't think our roommate will break our lease, but I am not certain", prompt the admission of graded beliefs. Indeed it seems plausible to suppose that beliefs to which we give full assent (probability of 1.0) should be reserved only for necessary truths. Necessary truths can be conceptualised as beliefs which have a probability of 1.0 in all logically possible worlds. Most people would grant that even deeply held views could after all turn out to be false. Hence most of what I am normally said to believe does not carry my full assent. So when we unpack the idea behind S's belief in P, we might find that S really means something like: S believes P *is likely*.

The adoption of this notion of belief provides a clear-cut route out of the Lottery Paradox. When belief 1 of the Paradox is unpacked, we find that it is actually shorthand for something like "I think it is likely that ticket 1 will lose but I grant there is a small chance it

will not." If this account is correct, then it follows that the remaining 99 beliefs regarding the lottery tickets are similarly qualified. Since there is a small chance that any particular belief about a ticket losing may turn out to be false, it cannot be concluded that no ticket will win.[9]

This solution can be similarly applied to the Preface Paradox. I only need to grant the qualified status of my belief that some of my beliefs are false in order to defuse the Preface Paradox. If what I really mean is something like: It is very likely that at least one of my beliefs is false, then there is no inconsistency in holding this belief. This resolves the Paradox in a much better way than Cherniak suggests because we preserve both our intuition that it is reasonable to think that one has false beliefs and that it is reasonable to be consistent. Of course this in no way undermines the possibility that there are some allowable inconsistencies. It is just that the belief-consistency paradoxes mentioned are not one of those cases. On the positive side, this suggestion should be welcomed since the sway of the binary view of belief may itself be based on assumptions from the Standard Model.

The idea here is that many of our beliefs have this probabilistic character about them. If so, then this may explain how people can come to have apparently inconsistent belief-sets. The Preface Paradox was one prominent instance in that regard.

The Prescription for Minimal Consistency

The normative requirements for Minimal Agents to maintain consistency should be nuanced with a distinction between recognising an inconsistency in one's belief set and having a obligation to resolve it. The discussion of the Preface and Lottery paradoxes was meant to illustrate that it may be better for the Minimal Rationality theorist to be somewhat

conservative in her acceptance of agent inconsistency. Perhaps we should only tentatively

accept that "inconsistency may sometimes be downright healthy" (Cherniak 1986a, 99), and

proceed instead on a case by case basis. Inconsistency may be necessary, perhaps even

healthy, but many of the cases in which it is so are distinguished by the peculiar property of

agent awareness.

Inconsistencies may be downright healthy only if they are hidden-nested amongst

beliefs stored in long term memory or obscured by the vagueness of the belief's content.

Cherniak's remarks about speed-computational time trade-offs and feasibility requirements

certainly support this view. Recall that if the cost of a memory search outweighs the benefits

of a successful search, an inconsistency may be preferred. Also, the discovery of some

inconsistencies may require more cognitive resources than the creature actually possesses.

Nevertheless, this is not to say that it also follows that inconsistencies which we are aware

of should also be similarly preferred.

The difference between inconsistencies we are not aware of from those we are aware

of essentially comes down to the differences between the properties of long and short term

memory. Presumably cognitive operations, including consistency checks, can only be

performed in short-term memory. Further, if it is the case that most persons can

automatically recognise inconsistencies in beliefs held in their short-term memory-up to the

seven item limit-then it may be that we should be required to always maintain short-term

consistency.[10] It might be suggested that if there is any universal rule the Minimal Agent

must follow it is "always resolve the inconsistencies you are aware of."

Feasibility constraints of the kind which made departures from formally correct

procedures allowable should not necessarily figure into questions about revising inconsistent beliefs we are aware of. Suppose I realise I am in contradiction by holding that p and that *not-p*. Surely anything I can come to believe I can come to disbelieve-even though it may be extremely difficult or damaging. So it should not be said that it is allowable to have inconsistency in *one's short term memory*. Still there may be some room for exceptions eleswhere, namely, with cases in which the rejection of the inconsistency would be too costly. Certainly, inconsistencies that arise from conflicting beliefs in long-term memory are sometimes allowable for the fact that memory limitations may make the recognition of such conflicts practically impossible, or, the tradeoffs between the inferential properties which determine the structure of memory make detection of inconsistencies between certain subsets of belief inappropriate.

Goldman (1986, 313-315) would take the further step of questioning the very need for consistency at all. He can think of only two reasons why consistency is thought to be harmful to an agent. First, an inconsistency is sometimes claimed to allow one to infer anything. Proofs in symbolic logic often rely on this method, for instance. But, as Goldman points out, we already know that people do not follow deductive closure. So the worry that persons having inconsistent beliefs will infer any and every logical consequence of their beliefs cannot be a realistic one.

The second possibility is that an inconsistency guarantees that one believes at least one falsehood. But an inconsistency logically guarantees only one falsehood and who is to say whether or not the harm of a particular falsehood outweighs any possible benefits (for the sake of argument). On the other hand, an inconsistency may result from many false

beliefs, so he grants that the value of consistency should not be dismissed altogether. He does also note that even a consistent set can have many (all!) false beliefs so it would seem to have no great advantage over an inconsistent set in that respect.

There is, however, a third possibility that Goldman overlooks. Inconsistency is an impediment to action. Whether or not one can make any sense out of the notion that an agent can both believe it is and is not raining, one surely cannot make sense out the notion that someone is going to both take and not take the umbrella. Having an inconsistent belief set can imply behavioural paralysis and that is why we should tend to avoid inconsistency. It should be granted that behavioural paralysis will not follow from inconsistencies between beliefs that have no real bearing on one's actions. What would really hurt are contradictions which involve objects and our environment: if you try to run and not run from the tiger, you get eaten. For this reason, one might be tempted to regard the irrationality of inconsistencies which follow from holding certain ephemeral beliefs as less worrisome: perhaps supernatural beliefs are one example. If so, panic over the widespread acceptance of irrational religious beliefs is unwarranted. Then again even ephemeral beliefs can have indirect effects. Consider how the supernatural (with other) beliefs of certain religious extremists have resulted in the murders of abortion-providers.

Cherniak's endorsement of inconsistency is to be supported only for contradictions which occur in long term memory.[11] Inconsistencies in short term memory should be regarded as much worse because of the threat they pose to an agent's ability to behave. It is conceivable, though, that there could be (weird) cases in which inconsistencies in short-term memory which do not threaten behavioural paralysis may be permissible.

73

2.5 Conclusion and Chapter Summary

The purpose of this chapter was twofold; first to expand on Cherniak's theory and second, to distinguish him from Stich. Stich's arguments against the epistemic virtue of truth were examined and addressed. Further support for the idea that true belief generating inferential systems were selected for was also given. The main trouble with Stich's comments here is that he merely *supposes* that our inferential systems were not selected for without giving any evidence either way to back that speculation up. Surely any particular gene or trait in a given species may or may not have been selected for. The question is why should we think that our brains evolved through some random process, as (perhaps) eye colour did, rather than by selection, as (say) our vision did. Jerrison's research on EQ suggests our brains did not develop randomly. Of course, even if selection played the dominant role in the evolution of the human brain it does not follow that our cognitive equipment tends to generate true beliefs. Optimal for survival may or may not be optimal for truth. Yet, I have argued that a trait need not be optimally realised in order to be valuable. Indeed, the position I have been defending has no need to claim that our beliefs optimise truth. I do make the supposition that brains that do not generate many true beliefs will be at a selective disadvantage. Correct behaviour requires true beliefs-although many harmless false beliefs (false positives) may also be allowed. I also grant that the connection between truth and selection is a conjecture, (albeit a reasonable one, I think).

It was also argued that the conclusions about truth could be generalised to apply to other features of inferential systems including cost, speed, power, memory size, and consistency. I will leave it up to the imagination of others to come up with more features.

74

There is no reason that the list of valuable qualities needs to be a short one. With the initial plausibility of epistemic virtues established, the second phase involved expanding on the notion. In particular, the main point here was to try to give convincing reasons to think that there are many features which must be mutually traded off against one another. The complexity of such interactions cannot be underestimated. The various environmental and hereditary forces behind the shaping of these tradeoffs must be staggering in their complexity.

Next I focused on one particular feature, namely, consistency. Consistency was singled out for greater attention mainly because I disagree with Cherniak's opinion of the ability of agents to maintain consistency. I would argue that he has underestimated the upper limit of consistency in reason-space. Although I agreed that global consistency was unlikely, Cherniak may have underestimated the feasibility of epistemic tradeoffs that yielded greater consistency. Depending on how the notion of belief is fashioned, the practical possibility for consistency fluctuates.

Finally, the intuitive resolution to the Lottery Paradox I proposed bears a close resemblance to Goldman's notion of probabilistic belief. If it is assumed that no beliefs-save a priori truths-are completely assented to by agents, then the ability to maintain consistency is greatly enhanced. This should pose no problems of feasibility since the increase in consistency is not the result of a strengthening of our estimation of our consistency checking mechanisms, but rather, a weakening of our notion of assent to (non-a priori) beliefs. To summarise, it is not inconsistent to believe that there is a twenty-percent chance that p is true and an 80-percent chance that not-p is true. If this approach to the Lottery Paradox and its

cousins is right then the upper limit of the Minimal Agent's capability to maintain consistency appears more sanguine than Cherniak has allowed. The moral: Cherniak's motivation to sketch upper limits to our cognitive capacities must be tempered with some caution. It would be a setback, and an insulting one, if we were to mistakenly underestimate the limits of the human mind.

Chapter Three: Minimal Rationality and Epistemic Amelioration

Introduction

In what follows I shall discuss the relationship between Cherniak's theory of Minimal Rationality (MR) and what shall be referred to as the 'Ameliorative Project'. With respect to epistemology, the aim of the Ameliorative Project is to improve our performance at gathering knowledge and making rational choices. It follows from the perspective of Minimal Rationality that the basis for that improvement should not necessarily be derived from the canons of logic and probability theory. What remains to be determined, then, is what the objective standard for rational inferential practices should be, if not logic. This claim should not be thought to be overly controversial, however, especially if one accepts that rationality is about desire-satisfaction in some fundamentally important sense. One simply needs to reject the view that rationality simply amounts to rigorously following the rules of logic and probability theory. Although it is clear that "*some* concern with logical and probabilistic relations is doubtless appropriate;...total preoccupation with logical or probabilistic relations is a misdirected employment of scarce attentional resources." (Goldman 1987, 369) Much remains to be said as to how the difficult path between formal and informal decision making is to be navigated.

In order to understand the place of the project of improvement within epistemology it might be helpful to think of the complete process as consisting of three stages or parts. The purpose of the first part, the *descriptive thesis*, is to describe what our inferential habits are. Here what is paramount are descriptions about what kinds of inferences we find easy, what kind are hard, what our biases are, and so forth.

There are at least two senses in which one could hold that normative epistemic questions are to be replaced by descriptive questions. Science can displace epistemology through either psychology or (for Stich) anthropology. But to think that normative epistemic questions really could be replaced by descriptive ones is to commit a naturalistic fallacy. Just because most people believe in God, it hardly follows that most people ought to believe in God or that most people must believe in God. Hence another difference between Stich and Cherniak is that Minimal Rationality begins with the descriptive thesis-which is determined by psychological and anthropological research-and moves on from there, whereas Stich is content to eschew prescriptive talk altogether. So there is still a need for a second phase of the Ameliorative Project.

The second part involves giving an account of what the normative epistemic ideal should be. In other words, what is needed is a kind of model which we should aspire to emulate. Note that such an ideal is distinct from the notion of degenerate idealisations which was discussed earlier. On my view, a degenerate idealisation was some theoretical aspect of a normative model which did not take actual human cognitive abilities into account. An epistemic ideal is simply a norm we should expect agents to live up to. Although the descriptive thesis is left to science to determine, this by no means implies that epistemology is to be left to the psychologists. The construction of a normative thesis would still remain even given a complete answer to descriptive questions. Kornblith (1985) criticises the view that psychological questions will ultimately replace epistemic ones, a view which he mistakenly attributes to Quine. This view is known as the 'Strong Replacement Thesis" and "on this view psychology replaces epistemology in much the same way that chemistry has

78

replaced alchemy." (Kornblith 1985, 6) In contrast is the Weak Replacement Thesis (WRT), or the view that psychological and epistemic investigations will eventually turn up the same results even if they are pursued independently of each other. This should be thought to be a good thing since epistemic discoveries should inform psychological ones and vice versa. Therefore even "if the weak replacement thesis is true, we can look forward to rapid progress in both psychology and epistemology as a result of their interaction, rather than either field being co-opted by the other." (Kornblith 1985, 8) Notice that the Weak Replacement Thesis is distinct from Psychologism. Even if it were true that the way we ought to reason is the way we, in fact, do reason, it does not follow that science and philosophy will eventually come into agreement. In fact, so far it seems that WRT is false: philosophers interested in knowledge and psychologists interested in human decision making seem to be involved in very different projects. The epistemologist is typically concerned with what is common to all instances of knowledge (say, justification, truth, and belief), while the psychologists tend to be more concerned with which cognitive processes are reliable. Both parties work is couched in different levels of abstraction.

One way of getting at the intuition behind the second part of the Ameliorative Project is through the performance/competence distinction. The descriptive thesis defines our performance while the normative model or ideal is fashioned based on our optimal physically possible capabilities. Note also that the normative ideal is the ultimate standard for rational agency and corresponds exactly with Cherniak's Strong Minimal Normative Ideal discussed earlier in this thesis. Since only a perfect agent could perform in accordance with the upper limit of the Minimal Normative Ideal, the likelihood of this actually occurring should not be

thought to be great. The chances of finding an agent whose performance did match the Minimal Ideal is analogous to the chances of finding someone who has never made a spelling mistake or multiplication error; it is physically possible but highly unlikely.[1] This is in contrast with the Minimal Descriptive limit below which attribution of rational agency cannot be made. In other words if you are a rational agent you must put in some minimal performance which meets the Minimal Descriptive limit. This still leaves room for a third and final phase of the Ameliorative Project.

The last phase involves giving feasible and appropriate suggestions towards moving from the descriptive thesis outlined by phase one towards the normative ideal described in phase two. The distinction between phases two and three correspond to the distinction between normative *principles* and *prescriptive* methods. Thus the complete plan involves explaining where we are, where we want to be, and prescribing how we should get there.

The preceding chapters have begun elaborating on what the normative ideal should be. What remains is to continue sketching a prescriptive account of how to move towards the Minimal Normative Ideal. It might be mistakenly thought that Cherniak has not yet provided a normative account at all, but this would be a misunderstanding. The view that the way we reason just is the way we ought to (Psychologism) is distinct from Cherniak's. Since agent performance as defined by Minimal Descriptive Rationality is thought to be always less than under the Minimal Normative Ideal, Cherniak maintains a prescriptive/descriptive distinction. The epistemic normative model we should aspire to just is Cherniak's Minimal Normative Ideal.

3.1 Chapter Goals

There are several main issues to be discussed, all of which bear on the issue of epistemic amelioration. The first shall focus on Putnam's (1981) treatment of the relativist/absolutist debate as it bears on the issue of epistemic justification as given in *Reason, Truth and History*. Putnam claims that epistemic absolutism-as characterised by Positivism-and Relativism are self-refuting doctrines. Cherniak's MR shall be distinguished from both of these undesirable positions. The second issue concerns the circularity of justification. There are reasons to worry that any kind of rational inquiry into rationality is self-refuting. How can it not be circular to use reason to evaluate reason? The third issue will look at three cases in which people would be described as acting irrationally on the Standard Model (SM) but not on an anthropological/relativist model. Examples from several different cultural groups, including the Azande, Kashdan peasants, and North American undergraduate students will illustrate certain points about good (and bad!) reasoning. Although the use of decision making procedures which deviate from formal logic may be preferable due to local opportunity-cost considerations, it seems that the differences between cultural groups (with respect to reasoning anyway) may have been overemphasised. Next, an analogy will be drawn between improving our learned abilities to make rational choices and other linguistic/rule following practices, including literacy and numeracy. I will suggest that the improvement of epistemic practices is similar to efforts to improve literacy rates or mathematical competence.[2] Indeed, the relationship between literacy, numeracy and rationality may be more intimate than mere analogy would suggest. Examples taken from Paulos (1990) will be used to suggest that the ability to make a rational decision is enhanced

when a certain competence in probability and statistics is present. In short, our inferential

practices could stand some improvement. The quality and effectiveness of any improvement,

however, depends on a combination of more research into non-formal heuristics and effective

training. Extensive knowledge of both our formal and informal tendencies could be used as

an aid towards narrowing the range of decision-making options which are most useful.

Research (philosophical and otherwise) may offer solutions to problems of adjudication

between strategies which offer different tradeoffs between reliability, speed, and so on.

Amelioration is made possible when these insights become well known to the general

population.

3.2 Some Lessons from the Philosophy of Science: Kuhn vs. Carnap

In *Reason, Truth and History*, Putnam compares the two most influential philosophies of

science in the Twentieth century, one of which is motivated by an absolutist picture of

rationality and is characterised best by Carnap and the Positivists, while the other is

motivated by a kind of epistemic relativism as developed by Kuhn and his followers, notably

Feyerabend. Putnam argues that the debate between the Positivists and Kuhnians has

overlooked the fact that both theories are self-refuting! Putnam thinks that both the tension

between these theories and the source of their incoherence stems from shared assumptions

about the primacy of science in epistemology-what Putnam calls 'scientism'. Roughly

speaking, scientism is the view that the answers to epistemic questions should be found in

science. For Positivists, an ideal epistemic model would resemble the methods of inquiry

as found in the hard sciences, i.e. physics, while for Relativists appropriate epistemic norms

are discovered by anthropology.

Putnam's arguments will be employed in order to show the undesirability of these positions. Although Putnam may dismiss Positivism too quickly, it shall be identified as a version of SM and is therefore open to the same objections. Epistemic Relativism on the other hand seems self-refuting in spite of Putnam's brash treatment. It must be stressed that Putnam maintains that his critique does not rule out objective standards for rationality, and likewise, epistemic norms. The existence of such objective norms, however, does not imply that there are necessarily ways to know with certainty what those standards are. I would add that the determination of norms is further complicated by context, culture, and individual differences. Although I find Putnam's remarks useful, I will take issue with his claim that epistemology would be better off with less science. Indeed, the position I have committed myself to in this respect is closer to Goldman, Kornblith-and of course-Cherniak, who argue that we should infuse *more* cognitive psychology into epistemology.

Positivistic Construals of Rationality

Within Logical Positivism, rationality was thought to be best explicated through various kinds of idealisations based on logic. Goldman cites the example of Carnap who wished to evaluate human cognitive performance through an ideal model-a strategy already popular in science:

> *Rational credence* is to be understood as the credence function of a completely rational person X; this is, of course, not any real person, but an imaginary, idealized person. (Carnap in Goldman 1986, 279)

Carnap and other Positivists hoped that a list or canon would be found which exhaustively described the scientific method and "since, according to the Positivists, the 'scientific

83

method' exhausts rationality itself, and testability by that method exhausts meaningfulness...the list or canon would determine what is and what is not a cognitively meaningful statement." (Putnam 1981, 105) Putnam claims that Positivist philosophers, such as Neurath and Carnap, favoured analyticity because they desired an "objective *uncontroversial* foundation for their arguments." (Putnam 1981, 109) In effect, a successful explication of the formal rules of language would result in rational institutionalisable norms and practices. Putnam points out that such a thesis is self-refuting *because this criterion itself is neither analytic nor empirically testable.* The argument against the positivist approach is thus: If it is true that the only rationally acceptable statements are those which can be verified according to the institutionalised rules of language, then that statement cannot be verified and cannot be rationally acceptable. Hence he argues "it is self refuting to *argue* for the position that (rationality) is identical with or properly contained in what the institutionalized norms of the culture determine to be instances of it." (Putnam 1981, 111) Yet Carnap was well aware of this objection and had a reply ready. Carnap argued that the verifiability principle was only a linguistic proposal and so neither analytic nor empirical. So it is not truth-functional, but, this does not mean it is meaningless. If it is a useful principle, then it still may be the case that it is rational to follow it. Putnam is quick to level charges of incoherence, perhaps a little too quick, nevertheless, his discussion of Positivism is sufficient to identify its Mark of Cain: SM assumptions and adherence to the Establishment view. The Positivist approach to epistemology can be criticised, for example, for failing to account for the limitations of actual agents. It is crucial to notice here, though, that there is no need to further conclude that rational argumentation and justification are

84

impossible. For Putnam, however, finding that justification is made more difficult when philosophers look towards science as a guide. For him "the way to develop a better understanding of the nature of rationality-the only way we know-is, likewise, to develop better *philosophical* conceptions of rationality." (Putnam 1981, 105 my emphasis) Carnap and Neurath's project fails because "arguing about the nature of rationality...is an activity that presupposes a notion of rational justification wider than the positivist notion ..." (Putnam 1981, 113)

Carnap's characterisation of Positivism is thought to be understood as a normative model. For Carnap, scientists should do science in the way he describes. But the history of science shows that real scientists do not do science in the way Carnap says they should. Although Carnap might insist that scientists should follow his model regardless of whether past practices fit his theory, the instrumental success of science might be thought of as a reductio argument against him. Given the success of practices which Carnap would count as mistaken, one is left wondering what makes his model preferable to what scientists already do? This kind of response to Carnap is developed by Kuhn who examines past scientific practices in great detail. Given Kuhn's historical approach, he may be thought to be engaged in a strictly descriptive project: not ascribing norms but rather merely taking account of how scientists actually do science. If this view were correct there could be no real argument between Kuhn and Carnap; one is engaged in a normative project, the other in a descriptive one. This view is mistaken. Since Kuhn acknowledges the instrumental success of science, his work can be taken as an endorsement of many of the past practices of scientists.

Kuhn and Relativistic conceptions of Rationality

Putnam next turns to the other major philosophy of science, a position motivated by Relativism, most cogently expressed in Kuhn's *The Structure of Scientific Revolutions* (SSR). It should be immediately noted that Putnam rejects any interpretation in which Kuhn is portrayed as rejecting all standards of rationality, although such a reading may be an accurate portrayal of more extreme relativists such as Feyerabend or Foucault. (Putnam, 1981, 114) Instead, Putnam focuses on Kuhn's thesis of incommensurability.

According to the thesis of incommensurability the terms used by another culture cannot be equated in meaning or reference with any terms we utilise. Putnam cites the example of electrons. Since Kuhn claims scientists in the early twentieth-century inhabited a different point of view-or if you prefer, paradigm, world, conceptual scheme, or gestalt switch-than we do, their notion of 'electron' cannot be equated with ours. Thus talk of comparing one style or method of scientific inquiry to another is nonsensical since there is no basis for such a comparison. Therefore there can be no objective standard which can be used to tell us which method of inquiry is most rational. Then again, it could be argued in Kuhn's defence that communication across paradigms is not impossible, but rather, partial and imperfect.

As with the Positivists, Putnam thinks the thesis of incommensurability results in another self-refuting conception of rationality. In reply to the Kuhnians, Putnam points out that if scientific theories from different times and cultures really were incommensurable (in a robust sense) then we could not make meaningful translations of theories from different worlds into our own. It would seem that to "tell us that Galileo had 'incommensurable' notions *and then go on to describe them at length* is totally incoherent." (Putnam 1981, 115)

Quite so.

Kuhn and Feyerabend reject any convergence in scientific knowledge since scientists operating under different paradigms are not talking about the same things-ontologically or otherwise. Putnam notes that Kuhn, unlike Feyerabend, does grant the instrumental success of science. Again however, it is incoherent to understand the notion of instrumental success without some degree of fixity of reference. As the instrumental success of our science mounts must we necessarily lose touch with older methods? It seems not. We could build a Viking longship and use Viking methods to cross the Atlantic. But should not the thesis of incommensurability entail that past instrumental success cannot be reproducible? Putnam argues that if we assent to regarding past selves, ancestors, and members of other cultures as persons, we must attribute to them certain shared references and concepts, for instance, such as what it means to be transported.

Putnam also notes that it is a truism that total relativism is incoherent. It is contradictory to hold the point of view that there is no point of view more justified than any other for "if any point of view is as good as any other, then why isn't *the point of view that relativism is false* as good as any other?" (Putnam 1981, 119) This does still leave a response open the relativist: 'Yes but when I say relativism is true, I mean true *for me*.' The rejoinder here is that if statements like 'X is true relative to person P' is itself true then there are absolute truths. On the other hand, a die-hard relativist must say that such statements are themselves relative but, as Putnam notes, what this would even mean is totally unclear.

Is Minimal Rationality Distinct from Positivism and Relativism?

Putnam speculates that both views (Positivism and Relativism) result from a kind of

scientism found within philosophy. The scientistic character of Positivism is overt: hard sciences, most importantly physics, are to serve as the paradigm for all other philosophical inquiries. The connection between Relativism and science becomes clear when one considers the position of the cultural relativist. The cultural relativist says that what is true, justified, and so forth amounts to the practices of one's local culture.[3] Hence determining those practices is scientific in the sense that it is an anthropological question as to which particular norms a given culture follows. The moral Putnam wishes to draw is that there is too much emphasis on science in epistemology. Putnam thinks that as long as philosophers assume a scientistic view, there will be a swing between relativistic and absolutistic fashions. It may be that Putnam feels that to bring the findings of science into philosophy is to commit a naturalistic fallacy.

Whether or not Putnam has characterised his targets accurately, his criticisms of the extreme forms of Relativism and Positivism seem sound. Contra Putnam, there are two reasons why we should still welcome the infusion of science into epistemology. First, the descriptive questions, that is, the scientific ones, do inform us about the normative ones insofar as science tells us what the contingent limits of our cognitive capacities are. This notion is sometimes called the 'psychological realism constraint'. Science discovers what the 'cannot' of the slogan "cannot implies ought not" *is* in the sense of discovering the limits of our computational capabilities and memory. Consider, for example, the possibility that the equation that correctly describes some presently unexplained aspect of the universe is too long to be uttered in one lifetime. Second, as pointed out by Kornblith, if one rejects scepticism, one is already committed to ballpark psychologism. If, in fact, we do know

88

(many) things, then the way we gather knowledge must be roughly (greatly) like the way we ought to. So the cognitive sciences are useful to epistemology after all.

Nevertheless Putnam's remarks about normative standards are encouraging. Although he rejects relativism, he points out that we cannot necessarily know decisively what the objective standards for rationality are. Similarly it can be the case that although science is an objective enterprise, not every scientific question has a determinable answer-some scientific questions may be indeterminate: i.e. "Why did the atom decay *now*...?" Other questions may have a determinate but context sensitive answer: "Why did John have a heart attack?"; "...similarly, holding that ethical inquiry is objective...is not the same thing as holding the silly position that there are no indeterminate cases at all." (Putnam 1981, 48) The same could be said for any value judgements including those which address questions about rational norms.

The main goal in this section is to establish that Minimal Rationality is not committed to either Positivism or Relativism-at least not as Putnam has characterised them. If MR is committed to either of these views, then epistemic amelioration seems problematic. If MR is relativistic, then improvement seems impossible. The danger lies in accepting the thesis of psychologism. As in ethics, if relativism is true then from what standpoint can one make any criticism of another's normative system? Meaningful criticism needs to be based on more than just my (or my fellows') opinions, it must also be *true* regardless of what anyone prefers to think. The Positivists, in contrast, do have an easy prescription for amelioration, one needs only to internalise the rules of formal logic and induction. As I have argued, however, this approach is inadequate. Literally equating logic with rationality is to court

89

disaster: just try to follow deductive closure if you do not believe me. Cherniak does not

give an explicit definition of rationality, but it is clear he avoids relativism. The notion of

'feasibility', for instance, is fashioned objectively in terms of computational complexity.

Also, he suggests that agents have an innate capacity to consider certain inferences as more

appropriate than others. Since some heuristics are innate, reason is independent of culture

to some extent. Nevertheless, there will be a great deal of flexibility in our ascriptions of

rational behaviour given the opacity and context-dependency of reason. Our normative ideal

of rationality just is the limit of our competence which is determined by a combination of

culture, inherited cognitive tendencies, and personal idiosyncrasies. In the last section, I shall

argue that the Positivistic answer to the issue of amelioration is partially correct, although

far too simple. Teaching awareness of both logic *and* our prior inferential biases, including

knowledge of the respective advantages and limitations of each, is probably the best place

to begin.

3.3 The Circularity of Justification

Any ameliorative project must ultimately face up to justifactory questions: why this way and

not that? Putnam's critique of Positivism and Relativism raised the problem of providing

an ultimate justification for rationality. Yet finding ultimate justification for rationality

seems an intractable problem. This problem is a methodological one: how are the grounds

of the debate decided without begging the question? Allow me to elaborate. The

assumptions behind the psychological literature produced in the Seventies followed the

Standard Model of rationality insofar as deviations from probability theory or deductive

inference rules were the standards used to indicate inferential error. But if one asked the obvious justificatory question "Why accept logic and probability theory as the normative standard?", the obvious answer would seem to be "They-meaning the psychologists and the scholars who formulated the Standard Model-are the experts." Still, one may be left wondering what else the experts can ultimately appeal to but pretheoretic *intuitions*? Stich (1990) mentions Nisbett's frustration with critics who demanded that he justify calling certain inferences 'bad' and others 'good'. Stich points out that without some principled basis to show that some subjects are reasoning badly, "the debate between Nisbett and his critics would degenerate into an exchange of raw intuitions..." (Stich 1990, 9) Notice that it was the degree to which our pretheoretic intuitions mesh with preferences which are (thought to be) rational that was to be tested for in the first place! Now a different set of experts-the philosophers-claim that what is and is not an error must be re-evaluated in light of non-formal decision making procedures. What rational basis is there to decide which account of rationality is the right one?

Can norms of rationality be defended without circularity? It may be unavoidable. However, Cherniak's remarks on the related problem of epistemic bootstrapping may be helpful here. Cherniak begins by asking how an agent can decide which of the infinite possible lines of inquiry are likely to be the most useful to an organism. The problem is that it seems to require 20/20 foresight to recognise what we do not know and what we ought to try to find out. Cherniak, following Plato, suggests that people "already have some inborn ("confused") knowledge..." (Cherniak 1986a, 117) He claims that there is some evidence to support this view. The structure of human (and all other vertebrate) vision follows a

hierarchy of attentional mechanisms. Our peripheral vision always detects motion rather than fine details, with only a very narrow central field dedicated to high resolution perception. Our natural reflexes will orient the eye towards motion detected by peripheral vision. Hence "the visual system thus embodies the assumption that what moves in certain ways is more likely to be interesting than what does not." (Cherniak 1986a, 118) A similar story can be told about other senses, such as the auditory system's startle reflex. Cherniak points out that these innate informational-selection strategies need not be optimal since "educated guesses that are better than chance can suffice" (Cherniak 1986a, 118) and many layers of learned routines as well as conscious decisions can supplement the innate processes.

The need for an innate inquiry-selection process becomes apparent when one realises that a regress threatens. A similar point can be made regarding why-questions: for any explanation, one can always ask the child's question "Why?" The regress of inquiries into previous inquiry-selection decisions would be infinite without postulation of an innate inquiry-selection procedure. The problem of bootstrapping, then, is raised when one considers how infants begin gathering empirical knowledge. An infant needs to be innately equipped before she can begin to accumulate beliefs. This hypotheses suggests that "evolution has simply constructed our heuristic preprocessing procedures so that we will ordinarily never even try to consider (the infinite number of counter-possibilities to any claim); we literally unthinkingly take their irrelevance for granted." (Cherniak 1986a, 118)

Of course, these facts only explain the norms we follow, they do not provide any ultimate epistemic justification. The point to be made here is that it simply may not be possible to provide any ultimate epistemic justification for our normative standards. I

suppose at some deep level, reason is innate-like breathing, we just do it. I must therefore assume the view that we can look no deeper than into our intuitions, and, most of those intuitions are based on the aforementioned innate preprocessing procedures. In the end justification can only be sought through an exploration of our innate intuitions, or, perhaps those of the experts. This conclusion need not force us to be gloomy, though, given Kornblith's remarks about ballpark psychologism. If we reject scepticism we are in business: the answers to descriptive questions about our inferential norms will also (partly) inform the questions regarding justification of our epistemic norms.

3.4 Are Formally Incorrect Decision Making Procedures Necessarily Irrational?

Whether formally incorrect decision making procedures are irrational or not depends on what is meant by '(ir)rational'. If it is supposed that the only rational inferences are those which are formally correct then the answer is-of course- "Yes." But it has already been argued that construals of rationality based solely on formally correct procedures are mistaken. Another, better, sense in which 'rationality' can be understood is through desire-satisfaction. Hence, on this view inferences, heuristics, procedures, and biases that tend to facilitate desire satisfaction are those which should be counted as rational. Irrational procedures are those which tend to frustrate desire satisfaction.[4] Given this interpretation it is clear that formally correct procedures will not necessarily be rational at all; for instance recall how speed/reliability tradeoffs could avoid computational paralysis. Even notwithstanding paralysis due to human computational limits it is easily imaginable that there are some informal methods that provide roughly reliable outputs at near-guesswork speeds (the

inferential equivalent of the speed of light). And indeed, on Minimal Rationality, any better-than-chance heuristic is counted as rational if there are no other feasible alternatives.

An advantage of this conception of rationality is that it avoids the circularity of justification. There is, in principle, a standard for determining which inferences will be counted as rational, namely, those which tend to satisfy desires. This notion is independent of our intuitions about logic and probability. Another advantage of this view is that it avoids any commitment to relativism as long as it is the case that certain inferential habits happen to be the most effective and feasible means by which to fulfil certain desires in any and all cultures. The utility of other strategies may vary depending on context-it is an empirical question. Nevertheless, the existence of common heuristic traditions in different cultures as well as the existence of certain innate heuristics would suggest that many of our inferential predilections should be universally accepted.

The justification of the formal procedures assumed to be rational on the Establishment View only follows if such procedures are valued intrinsically. But why should they be so valued? If rationality is about desire satisfactiont then only those procedures that tend to satisfy desires should be thought to be rational. This sort of view implies that there may be more than one way to get what you want, possibly even within the same socio-environmental context. Again it seems that a certain amount of vagueness is going to be necessary. It should also be stressed that it would be unsurprising (and welcome) if there should turn out to be a convergence between formally correct procedures and near optimal heuristics as Cherniak, like Goldman[5] and others would agree. In other words, it is only a contingent truth that logic is suited to rationality.

The value of a full description of our cognitive habits is to help us understand our biases. Why should anyone think that we must be prisoners of our inferential inclinations? What would be best would be to pick and choose the strategies, formal or otherwise, that tend to help satisfy our desires. So the criterion of a rational procedure is defined independent of rationality: the criterion is simply whatever helps satisfy our desires. The important thing to remember is that just because a given heuristic *is* more practically useful than a formal procedure does not mean that it is harmful to improve our knowledge of formal procedures. Learning logic does not undermine one's ability to make rational decisions.

Cherniak (1986a chapter four) showed that heuristics based on speed/reliability tradeoffs may be preferable to formally correct procedures. Furthermore, cannot-implies-ought-not considerations were raised in response to developments in complexity theory. These constraints follow from the highly contextual nature of rationality. I added previously that various other tradeoffs between various inferential virtues may result in a great many such informal decision-making procedures. Again, the main point is that any procedure is evaluated ultimately on the criterion of desire satisfaction.

In different contexts the degree towards which particular informal strategies are feasible and appropriate may vary. Exceptions to formal procedures may be allowable depending on the surrounding circumstances of the inference: the difficulty of making the computation, the costs and benefits of a given inference, and so on. In the next section I shall illustrate these points with anthropological data that fits this explanation.

3.5 Three Cultural Examples

95

The reasoning habits of three cultural groups shall now be briefly discussed in order to point out 1) their apparent differences and 2) their not-so-apparent similarities. I suggest that there are no inherent differences *in kind* between the reasoning of Azande, Kashdan peasants, and North American university students who were the subjects of some of the psychological research I have drawn upon. Although different cultural groups exhibit different particular biases, and all tend to display some difficulty reasoning logically, each of the groups mentioned demonstrates some ability to reason formally. I suggest that specific cases of systematic departures from formally correct procedures arise from either particular kinds of context-related factors which are determined by the local culture. In other words, some inferences are harder or easier to make depending on whether or not they are framed within a particular cultural context. I think that when these groups are compared, it is noticeable that they each share similar difficulties in their cultural patterns of reasoning. If so, there may be general methods of approach towards epistemic amelioration across disparate cultures.

Kashdan Inferential Habits

Consider first the Kashdan whom Halpike (1979) describes as incapable of completing certain simple syllogisms. For example, a Kashdan peasant was given the following information:

P1 There are no camels in Germany
P2 Berlin is in Germany

Next the subject was asked "Are there camels in Berlin?" The subject could not seem to conclude, even with prompting, that there are no camels in Berlin. This example may be taken as evidence in support of Cherniak's thought-experiment against 'Bridgehead'

96

arguments which assume that there must be some special class of inferences which "a subject must manifest if their mental states are to admit of any intentional description at all." (Stich 1990, 40) Earlier I mentioned Cherniak's example of an agent whose inferential feasibility ordering is very different from our own. The feasibility ordering of the Kashdan peasant would seem to resemble this description. Perhaps this example would be granted by Goldman as a successful prediction by MR.

Interestingly, the interview demonstrated that the Kashdan peasant could understand and use a basic syllogism as noted by Martin (1992). The peasant added the following during the interview: "If [Berlin] is a large city there should be camels there." and went on to conclude that "since there are large cities, then there should be camels. If [Berlin] is a large city, there will be Kazahks or Kirghiz [types of camels] there." Martin (1990, 95) points out that the peasant can reason logically since it follows from 1)There are camels in large cites and 2)Berlin is a large city, that 3)There are camels in Berlin. Martin speculates that the people in that culture have trouble reasoning about hypothetical situations. Other examples indicate that this phenomenon is not idiosyncratic to Kashdan peasants. The general conclusion that I wish to push towards is that there is an efficacy of real as opposed to hypothetical situations which functions as a facilitator of reason in other cultures, as well as our own.

Azande Witches

The reasoning habits of the Azande are sometimes claimed to be the paradigm example of a culture whose notion of rationality seems to diverge wildly from our own. The example of the Azande "supplies much of the grist to the relativist's mill and offers the most

97

obvious challenge to the rationalist, who is caught between the devil (at its most extreme) of logical relativism and the deep blue sea of writing off an entire culture as irrational." (de Costa et al. 1995, 326) The source of this distress are the Azandes' seemingly inconsistent beliefs regarding witchcraft. The Azande believe that post-mortem examination can determine whether a deceased person was a witch. They also believe that witchcraft is passed on by heredity to all same-sexed offspring. However, an Azande clan is a group of persons related biologically to one another through the male line. But the Azande do not infer from the premise that Man A of clan C is a Witch that every man in clan C is a witch.

As with the Kashdan peasant example, it would be a mistake to conclude that the Azande cannot reason logically elsewhere. In fact the data does indicate this. (de Costa et al. 1995, 327) Winch (1964) speculates that such an inference is not made because of the disastrous sociological implications. It seems that to be identified as a witch carries a social stigma normally reserved for the dead. Extending that stigma to all living members of a clan would create unneeded social strife. Also, since every clan has at least one male ancestor who is considered a witch, it logically follows that *every*one is a witch. Winch points out that if the Azande followed the principle of deductive closure here, there could be no distinction drawn between witches and non-witches. Hence, the whole enterprise of branding people as witches would be pointless. Finally, since beliefs about witchcraft play an important explanatory role in Azande culture-roughly equivalent to the role science plays in ours-it is convenient the Azande eschew making that inference.[6]

Two points from de Costa and French (1995) are important here. First, it seems that the Azande attempt to avoid having contradictory beliefs about witchcraft, and second, there

98

is a distinction between what Azande elites (namely, the ruling Princes) believe about witchcraft and what all other Azande think. De costa et al. argue that the Azande beliefs about witchcraft are *semi-propositional*. Such representations lack precise explication for the believer. The Azandes' supernatural beliefs are thought to be semi-propositional because when asked to elaborate on the content of those beliefs "people...wonder, argue or even fight about interpretations." (Sperber in de Costa et al. 1995, 332) The vagueness is advantageous insofar as it works to prevent individual Zande from realising the inconsistency of their beliefs. Hence they are not aware of any contradiction.

The Azande Princes, in contrast, are fully aware of the contradiction "since they know the outcome of every death in their provinces." (de Costa et al. 1995, 334) The situation is somewhat analogous to the difference between a lay-person's notion of electricity and a physicist's. Just as the lay-person in our society would go to an expert for an explanation about electricity, so to would a Zande appeal to a Prince for an explanation about the supernatural.[7] In both cases the lay-picture of the relevant phenomenon is incomplete and vague and the inconsistency is not noticed. This is not necessarily true for the experts: "Bohr was perfectly aware of the contradictory nature of his model of the atom." (de Costa et al. 1995, 335) Consider also the well known incompatibility between relativity theory and quantum mechanics which "breaks out into full-blown inconsistency" (de Costa et al. 1995, 333) at the foundational level. Likewise the Azande princes are aware of the deficiencies of their system, but unlike Bohr, they "do not search for a more consistent successor, since the current system serves their purposes." (de Costa et al. 1995, 335)

North American University Students

99

The last cultural group to be looked at are the subjects of the psychological literature produced by Tversky et al., Wason, Nisbett et al., and so on, namely, North American university students. Of particular interest here is the difficulty the undergraduates typically displayed when facing deductive inferential tasks. Wason found that performance on the selection task was worst when the problem was structured apart from a familiar context. Wason believed that when inferential tasks were presented in a non-formalised, familiar, setting performance improved. Wason's conclusions have been refuted by more recent findings by Cosmides. Cosmides discovered that unfamiliar problems which took the form of a social contract, like, "pay cost A in order to gain benefit B" were completed without the difficulties that Wason's Familiarity Theory predict. Cosmides appeals to an evolutionary explanation to account for humans' ability to correctly solve unfamiliar inferences formally identical to social exchanges. Other findings indicated that subjects could be mislead by the particular details of a problem even when those details were statistically meaningless.[8] Hence, like the Kashdan peasants and Azande, the undergraduate students displayed competence at some deductive tasks but not others.

Most members from all three cultural groups found difficulty performing particular kinds of formalised deductive tasks: the Kashdan with respect to hypothetical syllogisms, the Azande regarding the implications of their (largely unexplicated) supernatural beliefs, and the undergraduate students with respect to the Wason selection task. The Azande example demonstrated that inferential divergence from formal standards is not restricted to purely formal examples. Context can mislead as well as facilitate inference making.

It is important to note that the testing of university students did support the notion

that performance on all inferential tasks tended to improve with greater training in either statistics or formal logic. However, large proportions of those with formal training tended to make the same errors as the untrained undergrads lending to the conclusion that our inferential tendencies are very deeply ingrained.

Anthropological research done in the Sixties reinforced Kuhnian-like conceptions of rationality, namely that different cultures followed distinctive patterns of reasoning. Since each culture could function within its own style, arguments to the effect that any particular pattern was inherently superior seemed at best implausible and at worst racist. What these interpretations overlooked, however, were patterns common to all groups. It seems that the Azande Princes were aware of the implications of their belief system while lay-persons in Azande society were shielded from realising any inconsistency. Perhaps the importance of consistency in Azande culture has previously been downplayed by anthropologists because of the superficial appearance of wild reasoning-that is, having a feasibility ordering radically different from our own. Maybe past researchers assumed that since the Azandes' feasibility ordering differed so radically from our own, they must also value epistemic virtues, such as consistency, differently as well.

Cherniak, Goldman, Kornblith, and Stich would all agree that the formal errors we make do not automatically mean we are acting irrationally. Yet if these motivations are taken to the extreme we have a problem (i.e. anything goes). How do you navigate between the extremes of relativism and absolutism? The tentative answer was that circularity results when we attempt to justify reason on formal principles of logic or statistics; a better approach is to say that a reasonable strategy is one which helps facilitate the satisfaction of desires.

Yet this allows for the possibility that procedures which are completely at odds with logical principles may be counted as reasonable if they (accidentally) facilitate desire satisfaction.

There is a kind of tension in my remarks. On the one hand I argue that SM represents a flawed approach and that formal errors are not necessarily irrational. On the other it seems that not all departures from the laws of deduction should be allowable. For emphasis of this second point, consider this example taken from Tversky and Kahneman's (1981):

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternatives programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

Program A: (200 saved)

Program B: (600 saved, .33)

Most subjects chose program A. In a related experiment other subjects were presented with the same problem except that the options were:

Program A: (400 die)

Program B: (600 die,. 67)

Here the majority of subjects chose option B. What is surprising about these results is that the two versions of the problem are *identical*, except that the first options are posed in terms of potential gains, while the second are posed in terms of potential losses. The moral? "Subjects give different judgements, depending on whether outcomes are *described* as gains or losses." (Baron 1994, 365) What could be more irrational than contradicting one's own judgement just because a problem was phrased differently? In fact, the more recent

psychological studies suggest that the framing of a problem is *the* key to differences between making inductively correct and incorrect inferences (see footnote one of chapter one.) Like Kornblith, Baron speculates that such results may give us further hints about the inferential heuristics that people tend to use. Perhaps informal strategies such as "Avoid risks, but take risks to avoid losses", or, "Pay attention to outcomes that are certain", are used in place of formal procedures. Notice here, however, that it would be silly to invoke considerations due to feasibility constraints in order to justify heuristics which result in contradicting one's own judgement. It takes only a moment-if necessary with the aid of a pencil and (in my case) the back of an overdue power bill-to realise that these are simply two versions of the same problem. Hence it is not beyond our capabilities to overcome (this) bias. What is needed, though, is an understanding of elementary principles of probability and utility-expectation theory-and that is without question a matter of education.

The point, then, is that each group is capable of some improvement with respect to deductive procedures. Each could reason logically more often, but do not for various reasons. Furthermore, in all of the surveyed cases it is obvious that people would reason better if they had an awareness of their prior biases. What would help decision making would be greater knowledge of the reasons behind tendencies to solve problems in one way rather than another. So knowledge of formal procedures is not enough. The Azande society mentioned earlier would seem to require most people, and not just the Princes, having an understanding of the logical implications as well as the sociological motivations behind their system of supernatural beliefs.

3.6 Literacy, Numeracy and Rationality

In spite of the difficulties in providing an ultimate justification for rational norms, rationality is something objective that can be taught and learned. The situation will be shown to be somewhat analogous to improving literacy or numeracy: human reason is plastic so good reason can, to some extent, be taught.

Irrationality and Spooky Beliefs

The answer to the descriptive question (how we reason) tells us the limits to the normative one. If one assumes Ballpark Psychologism then descriptive inquiries are also helpful in determining how we should reason. If it is true that people know things "then many of their beliefs are arrived at by processes at least roughly like the processes by which they ought to arrive at their beliefs." (Kornblith 1985, 10) On the other hand Kornblith has already been criticised (my chapter 1) for being overly sanguine about our inferential tendencies. He notes only the cases of error (according to SM) that can be reinterpreted more favourably. But less sanguine descriptions are not hard to come by. Having irrational beliefs implies irrational inferences, or false data, behind the acquisition of those beliefs. False data, in turn, can be regarded as a kind of inference making from sense data and background theory. Consider now some widespread false beliefs.

I shall assume that supernatural beliefs are irrational, perhaps even paradigm cases of irrational belief, since the justification for such beliefs is so weak. By a 'supernatural' or 'spooky' belief I mean this in the ordinary sense, including belief in ghosts, angels, reincarnation, extra sensory perception, and so on. The interesting thing about spooky beliefs is that they persist in the general population (I will restrict myself to talking about North

104

America here) despite a general consensus of their fallaciousness among experts. Sometimes one runs across an attitude that Western culture represents the pinnacle of rationality. For those who find sympathy for such a view it may come as a surprise that 'spooky' beliefs are found across virtually all demographic borders. For example, recent polling data indicates that "nine in ten Americans think there is a heaven, seventy-nine percent believe in miracles and seventy-two percent in angels." (McAneny 1995) If these findings can be generalised to other nations, it seems there is cause for some concern if not panic. Belief in a wide range of spooky notions has actually increased over the past few years, including communication with the dead, reincarnation and belief in the devil. Only astrology[9] has lost ground although support is still at about one-quarter of the sampled population (about sixty-million people)!

Just in case it is presumed on account of some elitist bias that belief in the paranormal can be attributed solely to those who are poorly educated, it is instructive to note that even among college postgraduates spooky beliefs are still extremely popular. Although postgraduates were the most sceptical of any demographic grouping of those sampled, belief in the supernatural remained quite high with belief in God at seventy-five percent, Hell fifty-eight percent, and the devil forty-five percent!

Relativism seems to be of no help here since, after all, there is a minority of dissidents who do not believe in anything that is spooky. So, unless these people are different in some special way[10], presumably everyone could be convinced to believe that there is no God-and so forth-in the same way that people can be convinced to give up the idea that the earth is the centre of the universe. So there is hope after all. Memory limitations, speed-accuracy tradeoffs, and such, apply to everyone with human brains, but

assent to particular beliefs is never a physical necessity. This point, too, is in step with MR's distinction from the Fixed Bridgehead view.

Innumeracy and Irrationality

Other Gallup data indicates that irrational belief is not restricted to the realm of the supernatural. Misperception of basic facts, heightened by media hype and an inability to comprehend even elementary statistics results in some rather shocking misconceptions. For example, Moore (1995) found that in the aftermath of the bombing in Oklahoma city roughly half of those polled were either very worried or somewhat worried about either themselves or a family member becoming the victim of a terrorist attack. The misconception here is understandable from an emotional point of view, but still foolish. Paulos (1992) also remarks how a lack of understanding of probabilities can contribute to such misunderstandings. Since "if you don't have some feeling for probabilities, automobile accidents might seem a relatively minor problem of local travel, whereas being killed by terrorists might seem to be a major risk." (Paulos 1992, 9) He compares the seventeen Americans killed overseas by terrorists in 1985 with the twenty-eight million who travelled abroad (1 in 1.6 million) with the much greater chance of dying in a car crash (1 in 5,300). Hence activities which are most dangerous (smoking and driving) tend to receive relatively little attention. So "the natural tendency of the mass media to accentuate the anomalous, combined with an innumerate society's taste for such extremes, could conceivably have quite dire consequences." (Paulos 1992, 132) And it does.[11]

So if being rational is at all like being literate or numerate, then improvement is partly a matter of education. Presumably one can learn to make better probabilistic choices with

even an elementary understanding of the laws of chance. The statistics generated by Tversky and Kahneman certainly seem to support this conclusion. Those with an education in statistics or logic do tend to perform better-that is they make fewer formal errors-although it has been pointed out that the data is somewhat ambiguous. Goldman cites two studies, one which seemed to support the view that education did not improve performance very much and another which did. (Goldman 1986, 309) But it is hard to imagine that anyone would suggest that learning logic is going to do anyone any *harm*. Just because people generally get by quite well using heuristics hardly means they could not get by even better with a little formal training.

Certain biases tend to cloud our judgement in other areas of our lives, such as how we tend to think about motion and forces. According to Aristotelian physics heavy bodies fall faster than light ones. This principle is well supported through casual observation and works well as a rule of thumb, but is of course mistaken. Baron (1994, 92) cites research by Clement that found that students who had taken a physics course still held 'naive' theories of physics. Many students mistakenly thought that moving bodies required a force to keep them in motion, or that a ball shot through a curved metal tube would continue to follow a curved path when it exited the tube (figure 4), or that (figure 5) the path of a body moving off of a horizontal surface would not follow a parabolic path.
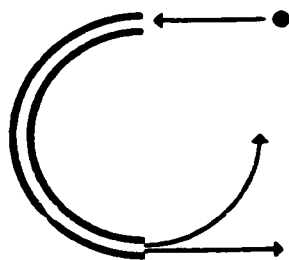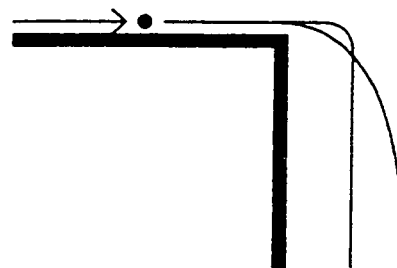
Fig. 4: Projectile Through Curved Tube   Fig. 5: Falling Body

Note that both figures are taken from Baron's (1994), pages 93-94.

Such misconceptions may be deeply ingrained in our thinking. Baron notes that the naive theories often corresponded with earlier theories proposed by early medieval scientists and Aristotle. Again the point here is that no one would say that tendencies towards such naive theories prevents people from learning physics.

The points raised by Baron and Paulos are meant to be taken together to help emphasise the idea that bias-awareness without some minimal formal knowledge of logic and probability theory is insufficient for amelioration (and vice versa). Having an awareness of the kinds of inferential patterns we are disposed to using is of no help if there is no complementary understanding of mathematical and statistical concepts. Again, this is not to say that all we need to do is fill introductory logic courses. The advantages that come from mastery of formal systems is not just prescriptive, but also, evaluative. Logic can assist in weeding out heuristics which lead to unnecessary contradictions, for instance. I envision an approach in which both formal and informal methods are taught in concert. In this way, particular reasoning strategies can be compared for both instrumental worth and deductive validity.

3.7 Conclusions and Summary

The discussion of Putnam's criticisms of Relativism and Positivism were meant to emphasise the unsuitability of those extreme construals of rationality. The right account, however, must borrow certain insights from both without falling completely into either camp. Cherniak's Minimal Rationality was argued to be consistent with such a middle position since 1) MR

is committed to objective norms and 2) MR is committed to an interpretation of rationality which is context-sensitive and opaque. As for whether it really is consistent to have it both ways, as it were, I shall leave for another time.

The second issue involved the circularity which seemed to threaten any attempt to justify rational norms. Two replies were given. First, perhaps Minimal Rationality avoids some of the circularity by replacing a conception of rationality based on following formal rules with one in which those decisions which are counted as rational are simply those which tend to satisfy desires. It is a matter of psychological, social, and statistical investigation as to whether any particular heuristic does indeed tend to be useful in fulfilling the wants of the agents who rely on it. The other point was that since MR is committed to the view that our cognitive habits are to some degree the result of natural selection, certain inferential tendencies and predilections towards inquiry selection, dispose us to certain kinds of intuitions. Ergo, Plato's answer to the bootstrapping problem is endorsed.

The third topic dealt with the rationality of non-formal judgements. Much has already been said about the need to recognise the rationality of formally incorrect procedures. Insofar as the Ameliorative Project is concerned, it would seem that we should continue studying our biases. Knowing our biases helps us recognise the nasty bits of some otherwise useful informal strategies.

The next issue concerned the habits of three cultural groups. The data obtained from the first group (the Kashdan peasants) appeared to give some support for Cherniak's argument against the Bridgehead view. Together, the habits of the members of these disparate cultures showed that contrary to what Relativists often claim, different cultures

109

share certain epistemic values, for instance, commitment to at least some formal inference making and maintenance of consistency in the belief-box. Insofar as amelioration is concerned, this uniformity is promising. It may mean that the general strategy of teaching both logic and an awareness of cultural and innate inferential predilections is a healthy alternative to the attitude that logic and non-formal heuristics are mutually exclusive. The examples of false beliefs and misconceptions about statistics and physics taken from Paulos, Baron, and the Gallup data emphasised how systematic departures from thinking rationally about the world can stem from both a paucity in knowledge of formal deductive and statistical systems as well as our innate predispositions.

The common thread running through these seemingly divergent issues is perhaps made more clear by way of an analogy. As with some cognitive biases or illusions, there are certain perceptual illusions to which all humans are innately predisposed. Consider, then, the famous Müller-Lyer illusion (figure 6), in which two equal lines appear to be different lengths:

Fig. 6: Müller-Lyer Illusion
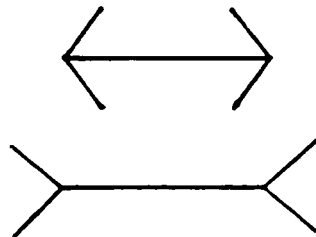
Occasionally, it is noted that knowing Müller-Lyer is an illusion does not make it go away. In another sense, however, the illusion does go away once one has become very familiar with it. One can recognise that the lines are in fact the same length even though our eyes tell us

otherwise. When I am asked "Which of these two lines is longer?", although my eyes tell me the lines are different lengths, I have seen the problem before, understand its cause (sort of) and answer correctly: "They are the same length." The solution to overcoming the perceptual illusion lies in our ability to distinguish between what our eyes tell us and the way the world actually is: we can simply take a ruler to confirm the supposition that the lines are equal.

Can some similar story be told with respect to other kinds of illusions, as with inferential fallacies and tendencies? Perhaps concise examples can be developed which help make the inferential biases we all have conspicuous in a way analogous to the role Müller-Lyer plays in teaching us about innate visual predilictions. The illusion is never fully overcome but we act properly once we learn to distinguish between reality and representation. The amelioration of reason requires much more than merely the rigorous application of formally correct techniques. Just as there is more to be learned from Müller-Lyer than proper usage of measuring sticks, there is more to good reasoning than simply avoiding formal error. Further criteria for improvement should include training to develop an awareness of our non-formal tendencies, that is, an understanding of the conditions which are likely to produce cognitive illusions. But beyond both of these suggestions, a kind of evaluative activity is also needed. Particular inferential habits, strategies, reasoning tricks, and so forth should be assessed with the intention of establishing their respective merits and deficiencies. In this way, near optimal, or at least the most feasibly optimal, strategies can be elucidated, compared and judged based on their value within a desire-satisfaction model of rationality.

Notes

Chapter One

1. See Holland et al. (1993) regarding deductive reasoning and Tversky and Kahneman (1993) regarding probabilistic reasoning. Other recent work by Cosmides (1989), Cosmides and Tooby (1996), Gigerenzer and Hug (1991), Fielder (1988), and Gigerenzer, Hoffrage, and Kleinbolting (1991) suggests that humans are, after all, good at inductive reasoning. It seems that inductive performance is best when tasks are presented in terms of relative frequencies rather than as single case probabilities. Nevertheless, so far the evidence of widespread systematic deductive errors has stood up.

2. They do not necessarily share a commitment to naturalism (Stich), nor do they agree that the way we ought to make inferences just is the way we in fact do. The philosophers I have mentioned might all share what Kornblith calls 'Ballpark Psychologism', that is, the view that the way we arrive at our beliefs is *roughly* like the way we ought to.

3. For an in depth explanation of heuristics in general see Tversky et al. (1993) and for discussion of the 'representedness' heuristic see Kornblith (1993). Basically, an inferential heuristic is a strategy or informal rule people tend to follow when they make certain kinds of decisions. The Representedness heuristic, for example, is based one what Kornblith calls the 'law of small numbers'. It seems that in contradiction with accepted standards of statistics, many people tend to believe they can draw reliable conclusions on the basis of very small samples. This pattern of reasoning can often lead to fallicious inferences, as with the case of racist generalisations, for example.

Kornblith argues that there is no error in this strategy if generalisations based on very small samples happen to be about members of a natural kind. Hence, I only need to see copper wire conduct electricity once to know that copper is a conductor. Presumably, the idea is that our evolutionary ancestors benefited from the law of small numbers given the abundance of natural kinds in their environment. Creatures that did not learn to fear all (say) sabre-toothed tigers after one or two tiger-attacks took Quine's suggestion and followed a pathetic but praiseworthy demise. Other heuristics include the 'availability', (Tversky et al. 1993), and 'social contract' heuristics (Cosmides 1989).

4. Hooker makes a few remarks on this shared tradition between both Empiricist and Rationalist thinkers. Reason was conceptualised as "an eternal, perfect structure which we somehow 'apprehend' and can then use, locating our imperfections elsewhere..." (Hooker 1994, 181)

5. I share with Cherniak the view that being rational is about getting what you want; Cherniak implicitly assumes this view in his discussions of belief-desire sets. The idea that the cash-value of rationality amounts to desire fulfilment is not without its critics. One common objection, for instance, raises the difficulty this model faces in adjudicating between

conflicting desires in the same agent. That being said, I do not think that the objections are finally persuasive. Unfortunately, space limitations preclude a full discussion of this conception of rationality. It will have to suffice to say that this interpretation of rationality is an assumption from which the discussion proceeds.

6. Just because folk-psychology allows us to make many useful predictions and explanations it does not mean that most people have a Theory of Mind. Goldman (1992) notes that there is great controversy over whether we should regard our everyday folk-psychology as a theory. Nevertheless, if our scientific or philosophical theories cannot do at least as good a job as our highly informal and non-explicated folk theories, it seems we have a problem. What is supposed to be the advantage of a scientific theory which is poorer at prediction than its folk-rival?

7. Note that some inferences that are intractable for a single agent may not be for a group of agents working together. In Carl Sagan's novel *Contact* a spaceship is constructed which is so complicated that no one individual understands all of its functions. Similar situations also occur in the real world. Some experiments in particle physics require several people to complete. It is not inconceivable that certain computer designs will be so complicated as to be beyond comprehension by any one individual.

8. Cherniak discusses computational limits at length in *Rationality and Anatomy* Cherniak (1986b).

9. 'Ideal' must be carefully distinguished from 'idealised'. It is possible to have a normative ideal which is not an idealisation in the sense I have been discussing idealisations (on the Establishment View). The difference is that it is possible, but not likely, for an actual agent to behave as described by Cherniak's notion of a Minimal Normative Ideal.

10. See Hooker 1994 (footnote 27) for why the explication of 'some' is difficult to provide without substantial philosophical and scientific research.

11. Stich and Laurence (1992, unpublished).

12. As an aside, it may be of interest to note the vast body of literature on Philosophical Scepticism. For example, consider the following list of famous philosophical examples and thought-experiments that bear a family resemblance to what Kripke calls 'The Wittgenstein Paradox':
*The Problem of Other Minds
*James's Squirrel
*Russell's Five Minute Hypothesis
*Russell's Chicken
*Gestalt Psychology's Duck-Rabbit and Face-Vase. (see *Philosophical Investigations*)
*The Brain in a Vat (for instance, Dennett: *Consciousness Explained*)
*Wisdom's Invisible Gardener

*Quine's Gavagai
*Pietroski (1995): Kepler/Schmepler example: elliptical orbits vs. *schm*elliptical orbits)
*Clarke Bus/Whale example
*Goodman's Green/Grue
*Descartes' illusion/dreaming/evil demon
*Inverted Colour Spectrum thought experiment

One feature these examples seem to share is that each turns on the tension between two competing explanations of the same phenomena. Typically, both explanations are consistent with the available facts but one is preferable to the other for some reason (perhaps because it is intuitively more plausible, fits better with background theories, or whatever). In other words, although we cannot prove it decisively, it is commonsense to say that we are not brains in vats. One exception is the Duck-Rabbit example. Here, as with the other examples, the two interpretations are inconsistent but neither the duck nor the rabbit could be said to be more intuitively obvious than the other.

Kripke briefly mentions the affinity between the Wittgenstein Paradox and Goodman's Green/Grue example. Interestingly, it appears that in this case the connection between Goodman's grue and the Wittgensteinian examples is quite intentional and explicit. (see Goodman *Fact, Fiction and Forecast* 3rd ed., Bobbs-Merrill, Indianapolis, 1973 Part I and Kripke 1984, 20.)

13. Actually he hints that there may be several other kinds as well but for his purposes (and mine) the degenerate/simplifying distinction will do.

Chapter Two

1. For instance see Cherniak (1986a, 66): "For human beings, some of the basic features of this structuring should be the result of natural selection, since they would have been helpful for the survival in any likely terrestrial environment...The rest of the individual's particular organizational scheme is learned, some of it as part of culture, but much of it as idiosyncratic and flexible cognitive habits based on past experience."

2. Gould also notes that biology textbooks typically illustrate evolution with examples of optimal adaptation.

3. Stich's own arguments against the evolution by natural selection of true-belief forming inferential systems may undercut the feasibility of pursuing many of the alternative reference fixing inferential strategies that he imagines we could replace truth with. If Mother Nature cannot necessarily be counted on to produce optimal designs why should we think we can?

4. This, of course, does not rule out strategies which are acquired by learning. I make the uncontroversial assumption that human behaviour is the result of an intricate interplay

between both environmental and hereditary factors.

5. Clarke (1996) also tackles this question.

6. Jacobson thinks that Stich is addressing intrinsic value here but he is not.

7. It should be obvious that consistency is distinct from truth, hence my reason for counting consistency as an epistemic virtue. A set of beliefs can be consistent without containing any true beliefs.

8. It should be noted that Hooker does not feel entirely comfortable with this characterisation of human memory because it is too simplistic.

9. One might still object that even if I do qualify my belief that each ticket will lose with a probabilistic reading of 'belief', I would still only believe that it is likely that someone will win and not that someone *will* win (100% certainty).

10. Ask a subject to remember some sequence of symbols, say, "A; H; L; not-P; not-W; not-H; not-S". Then ask them if they can judge whether the set is consistent. My guess is that for any string of simple beliefs, an adult can remember (in STM), they should be able to correctly judge the symbols for consistency. As to whether consistency checks are performed as easily when the subject is a child is less certain.

This view, however, is still far too simple. People commonly do not realise the logical implications of even very small sets of beliefs. Many logic puzzles, paradoxes, and brain-teasers depend on this well established fact. Consider Russell's 'Barber paradox': There is a barber who shaves only those men who do not shave themselves, and every such man. Does the barber shave himself? The implications of the Barber paradox are often not visible to the uninitiated, nevertheless, the puzzle can be stated quite succinctly.

11. Still it seems a little mysterious that there are some conditions that only become problems when you are aware of them. On my view, it may not be irrational to have inconsistent beliefs if you do not notice the contradiction. If only the same could be said for broken legs and flat tires!

Chapter Three

1. There is no reason to believe that it is not possible to increase our approximation of the Minimal Normative Ideal. Just as we can improve literacy rates and mathematical skills, education is likewise one of the best ways to decrease the incidence of formal errors in judgement.

2. The analogy will not be perfect. One problem is that the notion of literacy is relatively

well understood: one is literate if one can read and write. In contrast, I have argued that the notion of rationality is necessarily vague.

3. Another difficulty for the cultural relativist involves giving a plausible account of what (and who) determines what will count as a culture in the first place. Furthermore it seems there is no hard and fast distinction to be made anyway. Is it possible to belong to more than one culture or sub-culture? Consider the possibility of cultural identities based on region, socio-political identity, class, gender, or ethnicity. If this seems plausible, what principled basis can someone use to adjudicate between conflicts between the values of two cultures *when that same person belongs to both*? I would argue that such a conflict can only be resolved by appealing to a notion of value that goes beyond cultural norms. But that would make cultural relativism a pointless enterprise. One might also wonder if such an endeavour is self-refuting since different cultures may even disagree about the definition of 'culture'.

4. Martin (1992, 103) suggests that the criterion for rational behaviour should be truth and not usefulness. He points out that there could be a strategy of reasoning that results in a lot of false beliefs which happen to well serve the practical needs of the believer. Two replies here. First, Martin himself acknowledges the intuitive persuasion of the wish-fulfilment model: "Would you prefer not to think about things rationally? Would you prefer to act in ways that *don't* further your interests?" (Martin 1992, 36) Second, there are reasons to think that it can be a good thing to have false beliefs that are useful. Suppose Jane falsely believes that the bus coming down the street is actually a right whale cruising towards her. Although Jane knows that right whales are usually quite friendly to humans, she jumps out of the way anyway since it might not have noticed her. Here, it seems, Jane's false belief saved her life (like the 'Harry' example). Unlike the 'Harry' example, though, there was an aspect of her false belief that was true, namely, that a massive body was approaching rapidly. This true aspect of her false belief was relevant to the correct behaviour: jumping out of the way. Hence, insofar as Jane's desire to avoid becoming a pancake is concerned, it matters not whether she thought the object was a bus or a right whale. Notice that this could not be said if her belief had been false in certain other senses-say if she thought it was a cloud approaching, or, a Nerf bus.

5. See Goldman (1987, 369).

6. I am not suggesting that lay-persons in Azande society are aware of the danger lurking behind deductive closure, although de Costa et al. argue that this is not the case with respect to the ruling elite.

7. As further justification for this analogy, consider that most lay-persons would be at a loss to explain how their automobile engine or television set worked.

8. The well known 'Linda' experiment is but one example. See Goldman (1986, 308).

9. A recent broadcast on a local television-news program about the comet Hyakutake

identified an astronomer as a professor from the Department of *Astrology* at Saint Mary's University. This error, I think, belongs in the realm of literacy rather than rationality, although there would seem to be a tension (to put things mildly) contained within the notion of an Astrology lab.

10. They all studied philosophy?

11. What about a case in which irrational beliefs are useful? Bad things do not always follow from false or inconsistent beliefs and in fact sometimes very desirable things follow from having false beliefs. It would be preferred, of course, that people believe things that are true and that people behave altruistically. But suppose the altruism of a certain theist stems from her religious beliefs? Suppose we know such a person cannot be convinced that there are non-religious reasons to act morally. Would it be a good thing for such a person to be convinced that supernatural beliefs are irrational? Exploding even one false belief is not necessarily desirable; a lot of other (good) things may go with it.

Bibliography

Chapter One

Cherniak, C. (1981). "Feasible Inferences", *Philosophy of Science 48*: 248-268.

_____. (1983). "Rationality and the Structure of Human Memory" *Synthese 57*: 163-186.

_____. (1984). "Proto typicality and Deductive Reasoning." *Journal of Verbal Learning and Verbal Behaviour 23*: 625-642.

_____. (1986a). *Minimal Rationality*, Cambridge, Massachusetts: The MIT Press.

_____. (1986b). "Rationality and Anatomy" in *Naturalism and Rationality*, Buffalo: Prometheus Books.

Cosmides, L. (1989), "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason?", *Cognition 31*: 187-276.

Dennett, D. (1988), "The Turing Test: A Coffeehouse Conversation, *Reflections*", in *The Mind's I*, New York: Bantam Books.

Goldman, A. (1992). *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge, Massachusetts: The MIT Press.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.

Hooker, C.A. (1994). "Idealisation, Naturalism, and Rationality: Some Lessons From Minimal Rationality", *Synthese 99*: 181-231.

Kant, I. (1952). *Fundamental Principles of the Metaphysics of Morals*, New York: Longmans, Green, & Company.

Kornblith, H. (1993). "Our Native Inferential Tendencies", *Readings in Philosophy and Cognitive Science*, ed. A Goldman, Cambridge, Massachusetts: The M.I.T. Press.

_____. (1985). *Naturalizing Epistemology*, Cambridge, Massachusetts: The MIT Press.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*, Cambridge, Massachusetts: Harvard University Press.

Pietroski, P.; Rey, G. (1995). "When Other Things Aren't Equal: Saving Ceteris Paribus Laws From Vacuity" *British Journal for the Philosophy of Science 46(1)*: 81-110.

118

Pollock, J.L. (1987). "Epistemic Norms" *Synthese 71*: 61-95.

Quine, W. (1960). *Word and Object*, Cambridge, Massachusetts: The M.I.T. Press.

_____. (1969). *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Stich S.; Laurence (1992). "Intentionality and Naturalism" [unpublished]

Tversky, A.; D. Kahneman. (1993). "Probabilistic Reasoning", *Readings in Philosophy and Cognitive Science*, ed. A Goldman, Cambridge, Massachusetts: The M.I.T. Press.

Wittgenstein, L. (1994). *Philosophical Investigations*, trans. G.E.M. Anscombe, Oxford: Basil Blackwell.

Chapter Two

Bach, K. (1992). "Truth, Justification, and the American Way", *Pacific Philosophical Quarterly 73*: 16-30.

Clarke, M. (1990). "Epistemic Norms and Evolutionary Success", *Synthese 85*: 231-244.

_____. (1996). "Darwinian Algorithms and Indexical Representation", *Philosophy of Science 61*: 27-47.

Dawkins, R. (1986). *The Blind Watchmaker*, London: Penguin Books.

Gould, S.J. (1977). "The History of the Vertebrate Brain", in *Ever Since Darwin: Reflections in Natural History*, New York: W.W. Norton and Company.

_____. (1980). "The Panda's Thumb", in *The Panda's Thumb: More Reflections in Natural History*, New York: W.W. Norton and Company.

Goldman, A. (1986). "Epistemology and the New Connectionism", in *Naturalism and Rationality*, Buffalo: Prometheus Books.

Jacobson, S. (1992). "In Defense of Truth and Rationality", *Pacific Philosophical Quarterly 73*: 335-346.

Lycan, W. (1991). "Why We Should Care Whether Our Beliefs Are True", *Philosophy and Phenomenological Research Vol.LI*, No.1: 201-205.

Millikan, R. (1993). "Compare and Contrast Dretske, Fodor, and Millikan on Teleosemantics", in *White Queen Psychology and Other Essays for Alice*, Cambridge, Massachusetts: The M.I.T. Press.

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

Chapter Three

Baron, J. (1994). *Thinking and Deciding*, Cambridge: Cambridge University Press.

da Costa, N. and French, S. (1995). "Partial Structures and the Logic of Azande", *Journal of Philosophy 32*: 325-339.

Goldman, A.I. (1978). "Epistemics: The Regulative Theory of Cognition", in *Human Knowledge: Classical and Contemporary Approaches*, New York: Oxford University Press.

Halpike, C.R. (1979). *The Foundations of Primitive Thought*, Oxford: Oxford University Press.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, Chicago: The University of Chicago Press.

Martin, R. (1992). *There are Two Errors in the the Title of this Book*, Peterborough: Broadview Press.

McAneny, L. (1995). "It Was a Very Bad Year: Belief in Hell and the Devil on the Rise", in *The Gallup Poll Monthly*.

Moore, D.W. (1995). "Americans Brace Themselves Against More Terrorism", *The Gallup Poll Monthly*.

Paulos, J.A. (1990). *Innumeracy: Mathematical Illiteracy and its Consequences*, New York: Vintage Books.

Plato, (1977). *Meno*, in *Classics of Western Philosophy*, ed. S.J. Cahn, Indianapolis: Hackett Publishing Company.

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

_____. (1987). "Why Reason Can't Be Naturalized", *Human Knowledge: Classic and Contemporary Approaches*, ed. P. Moser; A Vandernat.

Winch, P. (1964). "Understanding a Primitive Society" *American Philosophical Quarterly*, *Vol 1, No.4*: 307-324.