The Normative Character of Interpretation and Mental Explanation.

By Paul Thorn
B.A.(Honors), SFU, 1996.

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR THE DEGREE OF
MASTER OF ARTS
in the
DEPARTMENT OF PHILOSOPHY

SIMON FRASER UNIVERSITY
August 1998

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-37645-1

Canada

Abstract:

This essay is devoted to the study of useful ways of thinking about the nature of interpretation, with particular attention being given to the so called normative character of mental explanation. My aim of illuminating the nature of interpretation will be accomplished by examining several views, some of which are common to both Donald Davidson and Daniel Dennett, concerning its unique characteristics as a method of prediction and explanation. Moreover, some of the views held by Davidson and Dennett will be adopted, elaborated, and defended. The conclusions of these philosophers do not, however, form an acceptable whole. Thus I will attempt to moderate some of their views. In particular, I will attempt to show up the defects of Davidson's view of the mental by defending the possibility some sort of psycho-physical reduction. Despite such philosophical pretensions, major parts of this essay will be devoted to sketching the foundations of a method for the interpretation of intentional behaviour which I take to embody the key features of our ordinary practice of interpretation. In particular, I will attempt to sketch the bases for a *method* of interpretation which is sensitive to the *methodological* considerations associated with the seemingly unique normative character of mental explanation. To this end, I will also investigate the question of how certain formal measures of coherence can be made to yield models for understanding the actual and possible bases of interpretation.

Acknowledgments

Among the people who helped me to complete this thesis the greatest amount of gratitude is due to Ray Jennings. Ray's helpfulness exceeded all reasonable expectations. In addition to volunteering much of his personal time to read and comment on drafts of the thesis, Ray offered his own creative insights, and was a constant source of encouragement.

Although almost all of the philosophers who I have studied under while at SFU have had an impact on my intellectual development, and, in turn, upon this essay, the following teachers were particularly influential in various ways: Kathleen Akins, Martin Hahn, Phil Hanson, and Bjorn Ramberg.

Many of the graduate students here at SFU assisted me in developing the ideas expressed in this thesis. For this reason gratitude is due to the following individuals who attended a series of talks I gave on material drawn from my thesis, and shared their ideas and helped to define my own: Michal Arciszewski, Rob Sinclair, and Judith Stapleton. Gratitude is also due to Martin Allen who shared my interest in the topic of *interpretation and paraconsistent logic*, and also shared his ideas and helped to define my own.

A final word of gratitude is due to Bryson Brown, who generously took time out of his holidays to read my thesis and appear in Burnaby to be my external examiner.

Dedication

For Amanda, who has given me love and support throughout my studies, and during the writing of this thesis.

Table of Contents:

Introduction:

An interpretation is an assignment of meanings to a linguistic product, or an assignment of beliefs, desires, and intentions to act (and possibly other *intentional states*) to an object for the purpose of explaining its behaviour. Interpretation, in turn, is the act or process of providing an interpretation. The two corresponding varieties of interpretation obviously intersect. In the case of interpreting an assertion we simultaneously attribute meanings to the linguistic products of an object and attribute intentional states to the object corresponding to what we take the object to intend to mean by its utterances. Obviously, we often rely on the type of expressions uttered by an individual and prior assumptions about the meanings of these expressions as the basis for attributing intentional states.

One normally considers an object to be an *intentional agent* only if its behaviour can be predicted by the attribution of causally efficacious representations which come in no fewer than three types: beliefs, desires, and intentions to act. In recent years some philosophers have come to believe that our interpretations of ourselves and others as intentional agents are arrived at by the use of an implicit folk theory. (I will also use the expression *intentional psychology* to refer to any form of explanation which functions by the attribution of intentional states.) One way in which the folk theory is thought to be unique is that its application (mental explanation) implicitly assumes the rationality of the objects of interpretation. Some philosophers, Donald Davidson among them, have argued that in virtue of the assumption of rationality implicit in mental explanation, the folk theory possesses certain very special 'normative' characteristics having the consequence that it is not

possible that it could be "nomologically reduced to a physical theory", and that mental types could not be "definitionally reduced" to physical types.(Davidson 1970, p.216)[1]

Some philosophers have taken the putative normative character of mental explanation as grounds for asserting the groundlessness of the folk theory. Stephen Stich, for example, has claimed that it is a consequence of the normative character of the folk theory that the attribution of intentional states is holistic, and thus, intentional psychology is an unfit paradigm for the study of cognition, since useful generalisations cannot be cast in terms of holistically identified states (or properties).(Stich 1983) In contrast to Stich, other philosophers, such as Davidson and Daniel Dennett, have argued that the folk theory is perfectly legitimate, and, moreover, in some sense secure from the threat of elimination in virtue of the normative characteristics of the theory.

This essay is devoted to the study of useful ways of thinking about the nature of our folk practice, with particular attention being given to the so called normative character of mental explanation. By way of illuminating the folk practice, I will linger a little over several views (some of which are common to both Davidson and Dennett) concerning its unique characteristics as a method of prediction and explanation. I will adopt, elaborate, and defend some of the views held by Davidson and Dennett. But without modification the conclusions of these philosophers do not

---

[1] It is clear that when Davidson uses the term *definitional reduction*, he has in mind reduction of one theory to another by means of the definition of the types of states or predicates of the coarser grained theory in the terms of the theory to which it is being reduced. When Davidson uses the term *nomological reduction*, I take it that he means reduction which is constituted by the deducibility the laws of the coarser grained theory from the laws of the theory to which it is being reduced through the application of a set of *bridge laws* which link a subset of the coarser grained theory's predicates to predicates of the theory to which it is being reduced.

form an acceptable whole. Thus, I will try to moderate some of their views. In particular, I will try to show up the defects of Davidson's view of the mental by defending the possibility some sort of psycho-physical reduction. (Some comments on the possibility of the elimination of the folk theory will also be provided.) Despite such *philosophical* pretensions, major parts of this essay will be devoted to sketching the foundations of a method for the interpretation of intentional behaviour which implicitly embodies the key elements of what I take to be our folk theory. In particular, I will attempt to sketch the basis for a *method* of interpretation which is sensitive to the *methodological* considerations associated with characteristic normativity of mental explanation. To this end, I also investigate how certain formal measures of coherence (which I do not presume to be, in their particularity, part of our implicit understanding of intentional psychology) can be made to yield models for understanding the actual and possible basis of interpretation.

Chapter one begins with an argument to the effect that our interpretive practices ought to be thought of as theory-driven. While the debate over this issue may seem to be (to the phlegmatic observer) somewhat frivolous, the issue itself (soberly understood) is important to the conclusions we reach regarding the prospects for intentional psychology, given recent and imminent developments in the field of neuro-science.

The idea that we attribute mental states to ourselves and others on the basis of a theory is generally taken to originate in Wilfred Sellars' paper *Empiricism and the Philosophy of Mind.*(1963) (As far as I know, Sellars was the first to make the idea explicit.) To a large degree Sellars' aim was to challenge philosophical dogmas about the genesis of the concepts of *thought* and *sense impression.* More

recently, the thesis suggested by Sellars tends to factor into contemporary analytic philosophy as a premiss in an argument in favor of the elimination of the folk theory, and in the defence of the claim that we humans do not in fact possess states of belief and desire, etc. The *general structure* of the argument is as follows:

Premiss One. The folk theory is correct if and only if it is vindicated by a physicalistic theory (through some form of psycho-physical reduction).

Premiss Two. The folk theory will not be vindicated by a physicalistic theory.

Conclusion. The folk theory is incorrect.

Eliminativists who proffer this form of argument take premiss one to hold on the assumption that folk psychology is a theory. While so called 'industrial strength realists', such as Jerry Fodor, *seem* willing to accept premiss one (and instead reject premiss two), others, and Davidson in particular, claim that, in a particular sense, premiss two is knowable *a priori* to be true, but claim (along with Dennett) that premiss one is not a version, or implication, of the thesis that folk psychology is a theory. (This characterisation of Davidson's view is liberal. I do not think, however, that the characterisation is outrageously inaccurate. I will clarify Davidson's view and contrast it with my own in chapter 4.)

My main interest in discussing the theoretical status of our folk practice is to get things right. I take this to amount to showing that premiss two is not knowable *a priori* to be true. Moreover, I will argue that the methodological considerations which dictate that we ought to find the individuals we interpret to be rational do not entail the irreducibility of the mental. Despite this disagreement with Davidson, my

sympathy with his view will be evident in my final confession: I believe that the variety of conditions commonly claimed by eliminativists to be necessary for the vindication of intentional psychology are too strong.

I will approach the issue of whether our folk practice is theory-driven by discussing what has been called the 'mental simulation debate'. Generally people on the other side of this debate, so called *simulation-theorists*, think that folk psychology is not a theory-driven practice. Simulation-theorists, thus, tend to argue against the conclusion reached by eliminativists by rejecting premiss one of the argument cited. In effect, simulation-theorists believe that our folk practice requires no physicalistic vindication, since, they believe, the practice is not theory-driven, and folk psychology is not a theory.

Chapter 1:

The Theory-Theory and the Problem of Irrational Belief.

§1.1 A Defence of the Theory-Theory:

Much intellectual energy has been expended in the last ten years or so on what has been called the mental simulation debate. The current interest in the topic seems to have been inspired by papers written by Jane Heal and Robert Gordon (both 1986), though a disagreement between Dennett and Stich[1] (both 1981) set the stage for the debate by delineating the two basic positions. On the one side there are the *theory-theorists*, philosophers and psychologists who believe that our interpretations of the behaviour of ourselves and others as intentional agents are arrived at by the use of an implicit folk theory. The theory-theorist holds that we make inferences about the intentional states of ourselves and others on the basis of a single theoretical framework, that is, the theory-theorist holds that (for the most part) we draw on a single body of generalisations for the sake of predicting the behaviour of ourselves and others. On the other side of the debate there are the *simulation-theorists*. These philosophers and psychologists think that our ability to attribute intentional states to others is based on a capacity to simulate, or empathise with, their mental attitudes. Though Robert Gordon is an exception, proponents of the simulation theory generally hold that we have non-inferential knowledge of our own intentional states.

The differences between the two camps can be characterised in many ways. We might characterise the difference as being over whether we should construe our

---

[1] One should note that Stich's views about the nature of intentional psychology have gone (and continue to go) through periodic transformations. Thus one should not expect to find much coherence in the variety of views attributed to him throughout this essay.

folk practice as grounded in an ability or in a body of knowledge. A useful way of drawing out the difference between the theory-theorists and the simulation-theorists is to take the debate to be over the question of whether our interpretative capacity is *theory-driven* or *process-driven*. While this is helpful, we must bear in mind that there are many theory-theorists who are also materialists, and cite materialist reasons for thinking that all cognitive behaviour is process-driven. However, we can make sense of the distinction by recognising that at least a sector of our interpretative behaviour bears the mark of being theory-driven.

For cogency, the theory-theory requires a conception of *theory-driven process* sufficiently robust so that (contrary to Simon Blackburn (1995)) the *mental simulation debate* does not collapse. The theory-theory must give grounds for construing our intentional attributions as theory-driven in something like as robust a sense as the theory-drivenness of the application of physical concepts.

Why suppose that our interpretative capacities are theory-driven? I will try to answer this question by stages. First, I will expose as faulty intuitions which suggest that our folk practice is not theory-driven. These intuitions derive from two sources. One is the idea that we have non-inferential introspective access to our own intentional states. The other is the idea that folk psychology cannot be theory, since it lacks features typical of scientific theories. Next, I will consider empirical evidence that gives us grounds for thinking that our interpretative capacities are *knowledge-based*. Finally, I will appropriate (and elaborate) some considerations expressed by George Botterill as to why this knowledge base should be considered a theory as opposed to just an assemblage of lore.

§1.1.2 A Rationalisation of the Argument from First-Person Authority:

A seemingly formidable obstacle to the theory-theory is the common opinion that we have non-inferential introspective access to our own intentional states. What is required in defence of the theory-theory, then, is some explanation of how this opinion could be wrong. Such an explanation plays the role of undermining a common misconception. The potential faultiness of the intuitions follows from the fact that at least some of our immediate perceptions are themselves the result of *theory-directed inferences*. The point, to use a Kantian idiom, is that theory-directed operations and inferences are performed on sensations prior to the *synthesis* of experience. On this view it is perception itself that is theory-laden. The lesson to be drawn from this point is as follows: although we might think that we possess non-theory-laden perceptions that we possess intentional states, it may be that such perceptions are theory-laden.[2]

§1.1.3 Theories Need not be Explicit:

The theory-theorist must admit that if folk psychology is a theory, it is quite unlike paradigmatic scientific theories. People have applied the putative theory for thousands of years without regarding it as one. It also appears that people who apply the theory typically have neither conscious access to the laws of the theory (as such) nor to the basis of their ability to apply the laws of the theory in particular situations. Given these facts, a defence of the theory-theory requires some sort of

---

[2] For a thorough account of the reasons for the potential faultiness of intuitions that suggest we have non-inferential introspective access to our own intentional states, see. Carruthers' *Simulation and the Self* (1996).

explanation of how folk psychology can lack the characteristics of paradigmatic scientific theories, yet still be a theory.

In defence of the peculiarity of the 'folk theory' several responses can be offered. I mention two. Both acknowledge the fact that we do not have full conscious access to the knowledge deployed in the practice of interpretation. The first response is simply to claim that while we do not in fact have conscious access to the folk theory, the theory is nevertheless explicitly represented in human brains and can thereby be applied in the process of interpretation in much the same way as more paradigmatic theories are first consulted and thereafter applied. Another way of defending the theory-theory is to acknowledge not only that the folk theory is not an explicit public artefact, but also that the theory is not explicitly represented at any unconscious level. The idea is that a systematised knowledge base is *tacitly represented* (at least in part) in human brains, and possibly in other conditions which shape the practice of interpretation.

The idea that the folk theory is tacitly represented in human brains is probably correct. However, the reason why we should think that folk psychology is a theory has to do with the apparent systematic basis of mental explanation. I will elaborate this point in section 1.1.5. For now I merely point out the following: that folk psychology is a theory is the best explanation of the systematic character of mental explanation. I take the conclusion that our interpretative practice must be grounded in systematising information-processing structures to follow from the fact that there is a corresponding systematicity implicit in mental explanation. Given this fact, would prefer not to argue for any particular conclusions regarding how the folk theory is represented. Rather I hold that the basis for our interpretative practice

must be grounded in *some* information-processing structures which systematise our means of understanding and predicting the behaviour of organisms such as human beings. Moreover, I take it that if we achieved a greater understanding of the physiology of the information processing structures which facilitate our interpretative capacities, we would come to appreciate that these structures *tacitly represent* our *implicit* folk theory.

I would now like to say a little more by way of clarifying what I mean when I say that a theory is *implicit* in our interpretative practice, and what I mean when I say that this theory is *tacitly represented.* When I say that a theory is implicit in our interpretative practice, I mean this: although a theory is never explicitly appealed to, the interpretations we tend to provide and the reasons one tends to give in justifying or criticising particular interpretations conform to the constraints that a theory would impose. Explaining what I mean when I say that I think that the folk theory is tacitly represented is more difficult, though I mean pretty much the same thing as Dennett.(1983) The general idea is this: a system *tacitly represents* some concept only if, one, the system is disposed to apply the concept correctly, and, two, states of the system can be interpreted as having semantic properties (relevant to the conceptual capacities of the system) in virtue of the globally defined functional properties of such states of the system. It would not, of course, be sufficient for theory-drivenness that the basis of our folk-practice be tacitly represented, though this would imply that the practice was, at least in part, knowledge-driven. Evidence that our folk practice is knowledge-driven when combined with other considerations lends support to the conclusion that the folk practice is driven by theory.

§1.1.4 Empirical Evidence for the Theory-theory:

Support for the claim that our interpretative capacities are knowledge-based has emerged from the research of cognitive psychologists. In particular, the performance of children aged around $3^1/_2$ at what has been called the *false belief task* seems to indicate that the difference between children who fail at the task, and those who succeed (generally humans 4 years and older) is a difference in knowledge and not mere ability.

In the developmental period between $2^1/_2$ and 5 years of age there are significant changes in the inferences children make about the mental states of other persons. These changes are demonstrated by the development of the capacity to succeed at the false belief task.

There are several versions of the false belief task, which typically test a child's ability to attribute mental states to others. In one version of the task, children are presented with a candy box, which is actually full of pencils.(Perner et al. 1987) The children are subsequently asked what another person will believe when that person is presented with the box.

> Three-year-old children consistently say that the other person will think there are pencils in the box. They apparently fail to understand that the other person's beliefs may be false. Again, this finding has proved to be strikingly robust. Children make this error in many different situations, involving many different kinds of objects and events. They continue to make the error when they actually see the other person respond to the box with surprise, and even when they are explicitly told about the other person's false belief (Moses & Flavell 1990)(Wellman 1990). Moreover, they make incorrect predictions about the other person's actions, which reflect their incorrect understanding of the other person's beliefs (Perner et al. 1987).**(Gopnik 1993)**

One might think that the results of experiments of the preceding kind, if they are not just neutral, actually lend credibility to the simulation-theory. The experiments seem to indicate that failure at the task is a result of an inflexibility in the child's ability to empathise. However, further experiments by Gopnik and Astington (1988) have produced striking results: children who fail on the preceding version of the false belief task also consistently fail at the task of retroactively attributing the correct beliefs to themselves. Children who fail on the standard false belief task also tend to attribute to themselves the belief, wrongly, *that there were pencils in the candy box* prior to having had the contents of the box revealed to them. This indicates that the inability to attribute false beliefs derives from a conceptual defect rather than from an inability to empathise. It should be mentioned that other studies (Wimmer & Hartl 1991) have shown that children of the same age can readily understand questions about the past; further, Gopnik and Astington's own studies indicate that children of that age can remember past events in the case of perceptual tasks.[3]

These and like results are generally taken to indicate that we attribute beliefs to ourselves via a psychological theory rather than by non-inferential introspection. This conclusion is no doubt too strong, since the supposition that the basis of our intentional interpretations of ourselves do not occur as a result of non-inferential introspection does not entail that the inferences to our own intentional states are theory-driven. Nevertheless, the evidence does lend strong support to the claim that our interpretative capacities are knowledge-based. Of course, both Blackburn and Gordon would argue that this knowledge base is quite slight and also that the

---

[3] For a more thorough analysis of such experimental data and a survey of the relevant literature, see Gopnik (1993).

predominant basis of our interpretative capacity is empathetic. This charge will be met in the following section. The important point which I take myself to have demonstrated here, however, is that there are grounds for thinking that our interpretative capacities are knowledge-based in at least some crucial respect. Furthermore, a point which must be recognised in defending the theory-theory is that it is not implied by the theory-theory that *no* part of our interpretative capacity is grounded in a process akin to cognitive simulation. The point is just that in interpreting intentional behaviour we employ theory-laden concepts and that our interpretations are constrained generally by theoretical generalisations.

§1.1.5 Why Our Folk Practice Ought to be Thought to of as Theory Driven:

A sound strategy in defending the theory-theory is one of drawing comparisons between scientific theories and the basis of our interpretative capacity. As Davies and Stone have pointed out, the theory-theory can be defended by pointing out some "*specific* parallels between the structure and the form of the explanations employed in folk psychology and the structure and form of explanations employed by science."(1995a, p.8)

One proposal for drawing out such a parallel is to adopt adequacy conditions which permit us to count behaviour as theory-driven only in cases where the behaviour is structured by information which is represented or used in a certain manner. Versions of this suggestion have been articulated by Botterill (1996), and also by Gopnik and Wellman (1992). In particular, in addition to recognising several

of the more standard characteristic of theories[4], George Botterill has argued that we should recognise that theories "must contain principles that provide a systematic integration of knowledge."(Botterill p.107) With reference to this demand, Botterill has nevertheless noticed that the frequently unrecognised characteristic of scientific theories admits of a degree of impreciseness, since cognitive economy, the practical benefit and the measure of systematic integration, is itself subject to degrees. Botterill's solution to the apparent problem is to point out that folk psychology is a paradigmatic example of a theory whose explanatory structure is integrated into a core of central principles. Here is a slightly revised version of the set of principles proposed by Botterill.[5]

The *action principle*: an agent, A, will act in such a way as to satisfy, or at least increase the likelihood of satisfying, her current strongest desire.

The *perception principle*: when an agent, A, attends to a situation, S, in a given way, and p is a perceptually salient fact about S, then A acquires the belief that p.

The *inference principle*: When an agent A possess the belief that *p*, and *q* follows logically from the conjunction of *p* with the other beliefs that A has, A comes to believe that *q*.

---

[4] These are, that theories make the provision for explanation and prediction (a necessary condition), that they contain principles of nomic generality (a necessary condition), that they postulate unobservables, and that they implicitly define concepts.

[5] It should be kept in mind that what is being argued for is the existence of a set of generalisations underlying our interpretive practices, and not that the particular set articulated here are the actual generalisations underlying our practice.

Botterill defends the principles which he suggests are constitutive of the core of folk psychology from the charges of their inadequacy (a charge which is made with reference to the obvious fact that none of the principles appears to be exceptionless), by arguing that exceptions to the core laws can be explained without embarrassment in the same way as exceptions to the core laws of more paradigmatic scientific theories. To this end, Botterill emphasises an observation of Lakatos': exceptions to the purported core laws of a theory do not show that the theory does not have such core laws, or that the proposed theory is not a theory, or that it is not a good theory.(Lakatos 1978) Rather, as Lakatos has claimed, it is typical for the core laws of scientific theories to have exceptions and for these instances of exception to be explained by appeal to a body of auxiliary hypotheses.

At this point I take myself to have established some good reasons for thinking that our folk practice is theory-driven. The general idea is that there seem to be generalisations implicit in our practice of mental explanation, and it appears that these generalisations and the concepts they implicitly define serve as a basis for systematising our knowledge of the patterns of behaviour which typically characterise certain types of organisms. The core laws (and the concepts they implicitly define) appear to serve as a basis for systematising our knowledge of human behaviour (non-intentionally described) by permitting representations of the fine grained-patterns of non-intentionally described human behaviour in terms of coarse-grained intentionally described patterns of behaviour. The behaviour of an agent is described intentionally by the attribution of contributory sub-states of the agent which exercise dispositions to influence the behaviour of the agent in virtue of their representational properties.

The conclusion that folk psychology is theory-driven is taken to be the best explanation of the systematicity of explanation implicit in our folk practice. If the form of reasoning which leads to this conclusion is cogent, the remaining difficulty lies is justifying the claim that there are, in fact, systematising generalisations implicit in our interpretative practice. It is, of course, particularly difficult to establish conclusive grounds for the claim that our interpretative practice is in accordance with some set of implicit laws. This is especially true, given the quantity of exceptions of which the suggested core laws admit. Despite this difficulty, I take it that many of the things I have yet to say about the nature of mental explanation will go some way in establishing that a theory is implicit in our interpretative practice. I will reiterate the relevance to the theory-theory of some of the things I have yet about the nature of mental explanation in the final summary of this essay.

§1.2 The Normative Character of the Intentional Psychology:

For some time now much has been made of an apparently distinctive feature of mental explanation: its projection of rational structure into the sets of intentional states we attribute to agents for the sake of explaining their behaviour. To my knowledge the primary proponents of this view are Donald Davidson and Daniel Dennett.

One of the central tenets of Davidson's philosophical psychology is that the principles of psychological explanation differ fundamentally from other kinds of explanation, in virtue of the fact that mental explanation is governed by a constitutive ideal of finding a rational order amidst the descriptions of mental events. Davidson takes adherence to this principle, that we find rationality in the descriptions of mental

events, to be exercised primarily by finding logical consistency in the content of the intentional states we attribute to agents.(Davidson 1983, p.316) Nevertheless it is also evident that he takes adherence to the constitutive ideal of rationality to be exhibited in taking the utterances of interlocutors to be responding to the same features of the environment as our own, in virtue of our judgments about the value (or necessity) of pursuing particular goals, and of recognising certain facts.(Davidson 1970, p.222) These principles of Davidson's view of mental state attribution are taken to be justified by an observation concerning how intentional states are identified: each intentional state is the state that it is in virtue of logical relations which obtain between the state itself and an accompanying cluster of states which together explain the behaviour of some agent. In turn, we must suppose, he claims, that a degree of logical consistency must obtain between any cluster of intentional states which together suffice to explain the behaviour of some agent. According to Davidson, "since beliefs are individuated by their logical properties; what is not largely consistent with many other beliefs cannot be considered a belief."(Davidson 1990, p.135) For Davidson the claim intentional psychology is irreducible and ineliminable is taken to be a consequence of the fact that one of the explanatory ideals of this mode of explanation is the projection of the rationality into the objects of interpretation, rather than attainment of *explanatory closure* which is arguably the ideal of the physical sciences.(Davidson 1970)

Dennett's understanding of the nature of folk psychology is similar to Davidson's. For Dennett the key features of intentional interpretation are that it is instrumentalistic and that the method incorporates a normative dimension. According to Dennett, folk psychology can best be viewed "as a rationalistic calculus

of interpretation and prediction, an idealistic, abstract, instrumentalist interpretation method that has evolved because it works and works because it has evolved."(Dennett 1991) Of central importance to Dennett is the defence of the supposition of the normative character of the folk practice. Dennett takes this supposition to support the claim that folk psychology is instrumentalistic inasmuch as the normative bases of interpretation depend on no picture of the physical realisation of the bases of behaviour.

Dennett's claim for an entailment from the normative character of mental explanation to the conclusion that the interpretation of an individual presupposes no picture of the physical realisation of the bases of his behaviour is to some to degree self serving. Dennett has acknowledged that "the strategic role [he] envisage[s] for the concept of intentional system" is that of "permitting the claim that human beings are genuine believers and desirers to survive almost any imaginable discoveries in cognitive and physiological psychology, thus making our status as moral agents well-nigh invulnerable to scientific disconfirmation."(Dennett 1980, p.73) Dennett's way of understanding of the normative character of interpretation is reflected in his proposed conceptual reduction of folk psychology.

In his paper "Three Kinds of Intentional Psychology", Dennett proposes a two-part reductive project, the first step of which is the 'conceptual' reduction of folk psychology (what Dennett considers a "vernacular social technology") to Intentional Systems Theory.(Dennett 1987, p.46) In Dennett's scheme, Intentional Systems Theory will borrow the terms of folk psychology, "belief", "desire", etc., and "give them a technical meaning within the theory."(Dennett 1987, p.58) Dennett describes the purpose of this reduction "as an attempt to prepare the folk theory for

subsequent incorporation into, or reduction to, the rest of science".(Dennett 1987, p.47)

Inasmuch as the proposed conceptual reduction of the folk theory to Intentional Systems Theory is supposed to be credible, Dennett proposes that there is a close similarity between the proposed normative science to which folk psychology is to be reduced (of which he takes decision theory and game theory to be fragments), and folk psychology itself. Dennett supports this claim by arguing that "We approach each other as *intentional systems*, that is, as entities whose behaviour can be predicted by the method of attributing beliefs, desires, and rational acumen according to the following rough and ready principles..."(Dennett 1987, p.49) The principles which Dennett articulates demand that we attribute the beliefs and desires that an individual *ought to have* "if it were *ideally* ensconced in its environmental niche"(Dennett 1987, p.49), and that we suppose that individuals act according to the dictates of rationality, given their beliefs and desires. After giving a brief account of the principles by which we attribute beliefs to intentional systems, Dennett sums up: "This gives us the notion of an ideal epistemic and cognitive operator or agent, relativised to a set of needs for survival and procreation and to the environment(s) in which its ancestors have evolved and to which it is adapted."(Dennett 1987, p.49) In turn, the effectiveness of the folk theory is explained with reference to a rough correspondence between how a rational agent ought to behave, and how agents who have been well designed by nature will behave.

Why construe our folk practice as a normative explanatory method? One reason follows from the observation that we implicitly assume the rationality of the

subjects whose behaviour we try to explain. Thus, in answer to the question, 'why did S to x?', we provide answers such as, 'because S desires O, and believed that by doing x, O would occur.' We need not add to such an explanation 'and S acts in rational accordance with her intentional states.' Another more complicated reason for viewing intentional psychology as normative reflects developments in twentieth century analytic philosophy.

Historically, the development of the view that mental explanation is normative seems to have come about in response to the apparent *principled failure* of analytic behaviourists to discover behavioural criteria for the attribution of various mental states. The premiss shared by people like Dennett and Davidson with the analytic behaviourists (of whom Gilbert Ryle is an exemplar) is that the evidential bases for mental explanation and the attribution of mental states has to be behavioural. How else could be attribute them? Thus when it became evident that behavioural analyses of mental concepts could not be given, philosophers attempted to articulate a new explanation for our ability to apply mental concepts on the basis of behavioural evidence. The new explanation emphasised the implicit assumption of rationality evident in our explanations of intentional behaviour. The general idea was that mental explanation is guided by a system of intentional concepts which aligns intentional state types into relations which are implicit in our understanding of what it was to behave rationality. For people who accepted this view, it was supposed that interpretation was a holistic process of attributing sets of intentional states that seem most appropriate as determinants of the physical behaviour of agents. Moreover, a principle condition for estimating the appropriateness of the attribution of a set of intentional states was that the attributed intentional states

cohered in the manner dictated by the relations which intentional states ought to stand to one another. Abiding by this principle condition for estimating the appropriateness of the attribution of a set of intentional states has generally been characterised as assuming the rationality of the object of interpretation.

As a matter of fact, I am more sympathetic to Davidson's characterisation of the normative character of folk psychology than Dennett's. In the first place, Davidson places less emphasis on what I will come to argue (in chapter two) are strategic reasons for assuming the rationality of the individuals we interpret. Moreover, while maintaining a claim for the normative character of mental explanation, the Davidsonian account has the virtue of being tolerant of the exhibition of less than ideal rationality when such irrationality takes the form of holding inconsistent beliefs. I will argue that this fact has the consequence that Davidson's account of the normative demands of interpretation is not susceptible to a criticism put forward by Stephen Stich against the aptness of the characterisation of our folk practice as a normative mode of explanation.

§1.3 The Problem of Irrational Belief:

Stephen Stich, opposing Dennett's proposed conceptual reduction of folk psychology to Intentional Systems Theory, has demanded a clarification of Dennett's view. Stich has asked Dennett to commit himself to the equation of rationality with a description of how we actually reason and behave, or with the deductive closure and logical consistency of an individual's intentional states. For obvious conceptual reasons (concerning the proposed normative character of the *intentional stance*) the first option is not open to Dennett, but the second option

seems also to be problematic, for presumably no *ideal system* ought to hold inconsistent beliefs.(Stich 1981, p.48) But in contrast to what must be seen as the most transparent dictate of rationality, Stich recognises the fact, as we all should, that human beings are often possessed of inconsistent beliefs. Given this fact, Stich concludes that we cannot follow Dennett and "agree to swap our folk notion of belief for the intentional system notion."(Stich 1981, p.48) Indeed, Stich thinks that the idealised normative Intentional Systems Theory cannot accommodate many of the descriptions of intentional behaviour which are possible for the 'folk psychologist'. The problem with the proposed reduction is, then, that the descriptive power of the new intentional theory will be impoverished and hence no proper substitute for folk psychology. The implication is that Dennett has mis-described the character of our folk practice, since the idealised version of the practice in the form of applied Intentional Systems Theory does not possess sufficient similarity to our folk practice to justify the characterisation of the folk practice as normatively-driven.

§1.3.1 Toward a Solution to the Problem:

As I have already argued, the most substantial ground for viewing our folk practice as theory-driven is the possibility of construing a set of three generalisations as representing the core of the theory which invisibly guides the folk practice. An interesting feature of these generalisations now assumes importance. The relationships between the intentional states types featured in the generalisations implicitly represent normative conceptual relations between the state types. Moreover, when an agent behaves in perfect accord with the generalisations her actions are perfectly rational (in at least one sense of what it is to be rational). It is

clear, however, that individuals seldom act in accordance with the standards which the core generalisations represent. Yet despite such regular exceptions to the core laws and the norms which they constitute, the principles still underlie intentional explanation in virtue of the presumption of the need to rationalise exceptions to the norms which the laws implicitly represent.

In attempting to *rationalise* exceptions to the core laws by appeal to auxiliary hypotheses there already exists an abundance of work which can be drawn from. We can find many examples of the application of auxiliary hypotheses in the canon of English literature, especially in detective novels where a protagonist typically attempts to deduce the motives that various characters would have for committing some crime. But in a purer philosophical form, we can draw upon the work of Donald Davidson who has argued that interpretative charity demands only that we not attribute *inexplicable* irrationality to intentional agents. He claims, "The methodological presumption of rationality does not make it impossible to attribute irrational thoughts to and actions to an agent, but it does impose a burden on such attributions."(Davidson 1975, p.159) For Davidson this dictate that we not attribute inexplicable irrationality is required to preserve the supposition of rationality inherent in the holistic constitution of the content of the intentional states of agents. Interestingly, Davidson's aprioristic conclusions comport well with what I would take to be the pervasive intuition of how one would go about protecting the core laws of intentional psychology in the face of their exceptions. That is to say, Davidson's conclusion is consistent with our actual practice. *Any inconsistency that we attribute to an interlocutor must be explicable, in principle, if we are going to have a legitimate basis on which to assert that we understand his meaning.* On my view,

the demands of preserving our folk theory include adhering to this requirement. As such, the dictate will be construed as an auxiliary law. (This law will be called *the auxiliary law of intentional psychology*, or simply *the auxiliary law*.)

Davidson has argued that irrational behaviour must be explained by 'rationalising' it. Rationalisations, according to Davidson, are fabricated to "enable us to see the events or attitudes as reasonable from the point of view of the agent."(Davidson 1982, p.289) It is clear that Davidson is correct in his diagnosis of the subjective character of acceptable rationalisations. A useful variation of the Davidsonian view leads us to the following thesis: failures of rationality are to be rationalised within the framework of *cognitive psychology*, construed broadly. Within this framework we can make a broad distinction between three types of cases. There are those in which we attribute irrational behaviour given a failure to attend to available perceptual inputs requisite for instantiating a particular belief (corresponding to failures of the *perception principle*), and there are cases where we attribute a failure to recall a premiss or to perform a suitable inference (corresponding to failures of the *inference principle*). For extreme cases of irrational behaviour, amounting to stark violations of the *action principle*, Davidson has resurrected aspects of a Freudian explanatory framework.(Davidson 1982) Davidson has claimed that we may tolerate interpretations of intentional behaviour which constitute attributions of inconsistent intentional states only by postulating a cognitive partition between them. According to Davidson "If we are going to explain irrationality at all, it seems that the mind must be partitioned into quasi-independent structures".(1982b, p.300) A similar cognitive structure to the one postulated by Davidson has been described by Christopher Cherniak. Cherniak recognises the

compartmentalisation of the mind as a significant factor in rationalising irrational behaviour, and, in turn, explains the possibility of inconsistent intentional states, in part, by postulating a distinction between working memory, and stored information. It is, of course, in working memory where we suppose that the content of intentional states may receive synchronic scrutiny. The appeal of the strategy employed by Davidson and Cherniak is evident: cognitive partitions are akin to the sort of cognitive rift which we habitually suppose to exist between different human beings *qua* cognitive agents.

§1.3.2 The Distinction Between Active and Inactive Intentional States and the Trouble with Contradictions:

According to Davidson it is a precept of interpretation that we must find coherence enough in the explanations of the behaviour of a putative agent to preserve what basis we might have had for seeing the behaviour as intentional.(Davidson 1973a, p.137) Davidson (as far as I know) has never attempted to make precise what the dictate that we find coherence enough demands. He has claimed, however, that we should never attribute a belief of the form α *&* $\sim\alpha$.(Davidson 1985, p.353) Satisfying this requirement is merely the most basic consequence of the auxiliary law. Yet it is clear that avoiding the attribution of such contradictions is not thought by Davidson to be sufficient for *finding coherence enough*. In any case, there are methodological grounds distinct from the ones offered by Davidson for the dictate that we try to find more coherence, rather than less, in the sets of intentional states we attribute for the sake of explaining and predicting the behaviour of agents. Consider the reasons why we should not attribute contradictory intentional states.

One respect in which attributing contradictory intentional states is problematic is very clear. Contradictory intentional states should not be considered the active determinants of behaviour, because, as is well known, contradictions permit any inference. Thus, the attribution of a contradictory intentional state can never be a non-trivial explanation of an action. Any intention to act would be an equally rational consequence of such a state. Moreover, if we were ever to suppose that an individual's behaviour was actively determined by a contradictory intentional state, any subsequent prediction of the agent's intentions to act would be equally licensed.

Given the supposition that an agent's intentional states are the causal determinants of his actions, we need a way of placing constraints on the intentional bases of an individual's actions on the supposition that his set of intentional states is inconsistent. One plausible solution to the problem is to suppose that an individual's entire set of intentional states is never simultaneously *active* as a determinant of behaviour. During times when elements of an individual's set of intentional states are not active we can think of them as *inactive* in virtue of the fact that despite not being causally efficacious determinants of behaviour while inactive, such states have the potential of becoming active at any time in response to changing circumstances. It is obvious that the attribution of some sort of regularity to the process by which sets of intentional states become active is important to interpretation and the prediction of behaviour generally. As a temporary measure, I will assume that, in general, the most appropriate or relevant intentional states tend to become active according to circumstance. Undoubtedly, the possibility of the activation of inappropriate intentional states is a potential source of irrational behaviour.

Setting aside the problem of saying how a set of *potentially active intentional states* become *active*, we must recognise some distinction such as the one proposed here between *active* and *inactive states*, since on occasion we find ourselves having to attribute inconsistent intentional states. The existence of inconsistent intentional states and the problematic nature of the consequences of such states as bases for the prediction of behaviour force us to regard only consistent subsets of an agent's set of intentional states as possible synchronic determinants of the agent's behaviour.

Now despite what I have claimed, there are examples that appear to show that we can attribute a contradictory intentional state as an appropriate explanation of an agent's behaviour. One type of contradictory belief is the belief that a statement α (in some formal language) is a theorem, when, in fact, it is not. Moreover, one sometimes finds oneself explaining the behaviour of a logician attempting to prove α (a non-theorem) to be a theorem, by saying that the logician believes that α is a theorem.

One point, which appears to soften the bite of the present example is this: if we suppose that rather than believing that α is a theorem, the logician believed merely that the sentence "α" expresses a theorem, then it seems that we could explain his behaviour without committing ourselves to the attribution of a contradictory belief. This way of avoiding the apparent problem amounts to adopting the use of quotation as a means to distinguishing ways of believing. Such a distinction is not unproblematic. For one, it is appears that an agent's believing a

sentence true will influence his behaviour in ways precisely analogous to a belief in the corresponding proposition.

Providing a wholly satisfying solution to the class of problem that has been mentioned lies beyond the scope of my present aims. As a temporary measure, I propose to stipulate the problem out of existence (or at least out of sight) by deciding to say that individuals will act according to all of the consequences of their *active* intentional states. That is, I will suppose that individuals will behave as though the actual world is among the set of models which satisfy their active intentional states. Given this deliberately heavy handed assumption, it would be intolerable to attribute a contradictory intentional state.

The invocation of this heavy handed assumption in intended to serve a practical purpose, although the underlying problem is conceptual. Assuming that an agent's intentional states are inconsistent we must circumscribe the *classical* consequences of these states from which we will make our predictions of the agent's action. The natural way to do this is postulate partitions between the inconsistent states (allowing only states not separated by a partition to be active at any given moment). There is no straightforward way of taking such a measure in the case of contradictory states. We must simply suppose that an agent who apparently acts upon a contradictory state is acting upon some *subset* of the unchecked set of consequences of the state. In this way, the attribution of a contradiction is never wholly appropriate. Why not just attribute a less implicationally promiscuous intentional state capable of explaining the agent's actions? Not only can there never be a case where an agent simultaneously acts according to all of the consequences of a contradictory state, there would be no

principled way of deciding how an agent would act assuming that his actions were determined by a contradictory intentional state.

With regard to such troubling examples as the one mentioned, the following strategy is proposed: if we have reason to attribute subscription to a contradictory statement, then rather than attribute the contradictory statement, we must attribute the belief that the sentence which expresses the contradictory statement is true. We may also attribute whatever other non-contradictory beliefs are capable of explaining the actions of the agent we take to follow from his mistaken belief that the contradictory sentence is true.

Far from showing that it would be acceptable to attribute a contradictory intentional state as a basis of action, the example mentioned does not go against the following point: if we attribute a contradictory intentional state, we will be obligated to place severe restrictions on our predictions of how the agent will act and what the agent will infer on its basis. Moreover, the general point of the considerations I raised against the practice of attributing contradictory intentional states stands. In interpreting the behaviour of an agent we attribute intentional states which explain his actions. In virtue of the logical properties of contradictory statements, the attribution of a contradiction as an active determinant of behaviour is capable of explaining any action, and thus such attributions are always explanatory, and are for this reason generally defective. A similar moral applies to the attribution of inconsistent sets of intentional states, since no inconsistent set of intentional states should ever be taken simultaneously to be a determinant of an agent's actions. Moreover, if we adopt the convention of considering a set's coherence as a measure of its resistance to the derivation of contradictions, the following

generalisation appears to follow. The more inconsistent an agent's set of intentional states, the more likely that it will be deficient as a basis for making predictions of the agent's behaviour, since an inconsistent set of intentional states is bound to dictate inconsistent courses of action. (This point will be defended in greater detail in section 2.1.5.)

In the light of these remarks, we can see that the attribution of extreme inconsistencies in the explanation of behaviour is simply *ad hoc*, and has the same deficiencies as any *ad hoc* explanation or theory. Moreover, the dictate that we maximise the coherence of the sets of intentional states we attribute is analogous to a dictate of parsimony. Given the preceding observations concerning the apparent relationship between the coherence of a set of intentional states and its *projectability* (that is, its integrity as a basis for making predictions of the behaviour of its possessor), we ought to recognise as a principle of the methodology of interpretation that we maximise the coherence of the sets of intentional states we attribute. I will refer to the methodological dictate as the *coherence principle*.[6] Here I have in mind the conception of methodology (similar to Cummins') as a system of norms for evaluating applications of a mode of explanation.(Cummins 1983, p.*vi*)

§1.3.3 An Apparent Solution to the Problem of Irrational Belief:

The solution to the problem suggested by the characterisation of intentional psychology given so far consists is as follows: we should supply quantifiable weakenings of the ideal of perfect rationality of such a kind that we can articulate

---

[6] In virtue of its function, the coherence principle need not be tacitly represented in human brains, since the impact of the principle inevitably constrains interpretation.

agent-relative guidelines for the attribution of intentional states. The demand placed upon our estimation of how an individual will behave, given his possession of particular intentional states, is simply weakened in proportion to our estimation of his deductive abilities. Moreover, it is clear that we could provide agent relative formalisations of the implications codified by the core laws of intentional psychology which would permit the characterisation of imperfect cognitive agents, and the prediction of their behaviour. One way to do this would be to define ideal rationality as logical consistency and comportment with an inferential procedure fixed at optimal values. This is tantamount to defining ideal rationality as genuine comportment with the dictates of the core laws of intentional psychology. Having done this, we could, by varying the values of the proposed inferential procedure, and by defining varying degrees of consistency, characterise less than ideal cognitive agents. In addition to the fact that all cognitive agents could be described as employing variations of the same inferential procedures, the same core principles and implicitly defined intentional concepts would be applicable to all cognitive agents. Thus, despite our development of models which permit the instantiation of inconsistent sets of intentional states, we could still maintain that the propositions which compose these sets have the same semantics. By this I mean that we may attribute to Smith the belief that not α, and to Jones the belief that not α and the belief that α, yet maintain that the content of Smith's and Jones' beliefs that not α are identical, and that Smith and Jones are possessed of type-identical intentional states.

§1.3.4 The Machine Model of Deductive Reasoning:

In seeking a suitable model with which to represent variations in deductive ability, one is naturally drawn to examine algorithms which are not unlike the ones realised by automated theorem-provers. According to what I will call *the machine model of deductive reasoning*, the basic types of cognitive attributes which we should use to make sense of agents which possess varying degrees of rational acumen are *inference preference-orderings*, *inferential conscientiousness*, and the *complexity of content for an agent*. Used jointly, these attributes permit us to make sense of sub-optimal deductive ability by simulating those features which are commonly accepted as reflecting the difficulty of a deduction. According to the accepted wisdom, the difficulty of a deduction is a function of the length of the derivation required to perform it, and of the difficulty of applying the rules required by the proposed derivation.(Johnson-Laird 1993 p.7)

The notion of *complexity of content for an agent* is probably the most fundamental concept for the present attempt to implement our folk theory in a manner sensitive to variations in cognitive abilities. It must be supposed that the contents of intentional states are in some degree and manner compositional, and that each agent possesses a relative complexity rating for each type of semantic primitive of which his intentional states are composed. Once we have assigned complexity ratings to the semantic tokens which an agent uses, we can assign a complexity limit to his working memory. Roughly, the working memory is the central processing unit of an agent. Working memory is the source of premisses for inferences and reasons for actions according to what intentional states are present.

The *complexity limit* of an agent's working memory measures the maximum amount of content which an agent can act upon at a time.

The distinction between working memory and storage permits us coherently to describe agents possessed of inconsistent belief sets without forsaking the core principles of our theory of rational agency. We can also make sense of the avowal of contradictory beliefs, since we can ensure ourselves that the content of a contradictory statement expressed verbally by an agent exceeds the complexity limit of his working memory. Given the assignment of complexity ratings to semantic primitives of an agent, and given the complexity limit of the agent, one can easily assure oneself that no agent possesses a self-conscious inconsistency, simply by adjusting one's assessment of the complexity ratings of the agent.

The *inference preference-ordering* and the *conscientiousness of an agent* determine the means by which the agent may draw inferences from his beliefs so as to make its 'implicit' knowledge 'explicit', and, moreover, eliminate inconsistencies from his belief set. The conscientiousness of the agent will generally be a value which determines the frequency at which the agent makes inferences from the subsets of his belief set so as to draw consequences from the information he stores. The inference preference-ordering of an agent simply describes the inferences the agent will make when particular structured content types are brought into working memory. A significant point about such inference-orderings is that not all of the inferences that an agent makes need instantiate valid forms of reasoning. Though an agent would be expected to make mostly valid inferences, some invalid but practical strategies might also be used.

§1.3.5 Dennett's Willful Oversight:

In his paper *Making Sense of Ourselves* Dennett argued that ideal rationality is not to be equated with deductive closure and/or logical consistency. Rather, he wishes to use the term "rationality" as a "general-purpose term of cognitive approval--which requires maintaining only conditional and revisable allegiances between rationality, so considered, and the proposed (or even universally acclaimed) methods of getting ahead cognitively, in the world."(Dennett 1987, p.98) This response to Stich differs strikingly from the one I have suggested. It seems, then, that Dennett has forsaken one apparent means to what he desires, and in particular a characterisation of folk psychology as a normatively-driven (and, hence, supposedly 'insulated') theoretical stance. Dennett's characterisation of rationality is also unintuitive, since it seems to entail the claim that what counts as a rational strategy is whatever it is that will best achieve the goals one has, given the environment one lives in. Dennett's characterisation of rationality is, thus, not in accord with the demands for intelligibility dictated by the core laws, and our intentional concepts. The types of irrationality that we, as interpreters, can make sense of are determined relatively to an agent's epistemic situation. Broadly speaking, an agent's epistemic situation consists in the agent's current beliefs and desires, his means of drawing consequences from these intentional states (or acting upon them), and his way of gaining knowledge about his environment. Moreover, the type of rationality which the core laws (and the associated intentional concepts) invite us to discover are also relative to the epistemic situations of agents.

In addition to the discordance of Dennett's conception of rationality with the character of our intentional concepts, the application of his concept seems to me an

unwelcome burden. Moreover, it seems that the conception of intentional psychology I have articulated (which includes *the auxiliary law* and *the coherence principle*) can ground a response to Stich. This response resonates with Dennett's articulation in *True Believers*(1981) of the claim that the attribution of intentional states requires the attribution of rationality. In *True Believers* Dennett writes that "One starts with the ideal of perfect rationality and revises downward as circumstances dictate. That is, one starts with the assumption that people believe all the implications of their beliefs and believe no contradictory pairs of beliefs."(Dennett 1987, p.17) Some such as this would be my response to Stich. In conformity with our folk theory we presume that agents possess consistent sets of intentional states, and that they will behave in accord with the core principles of our theory. However, whenever agents fail to behave in accord with the dictates of the theory (though it is quite possible that one prejudicially anticipates that most humans tend to behave irrationally in characteristic ways) auxiliary hypotheses are available to explain the anomalies. These rationalisations are dictated by the auxiliary law, the application of which is guided by estimating the reliability of the agent's belief-generating processes and the conscientiousness with which the agent applies his cognitive powers. Finally, we explain that Intentional Systems Theory will function by essentially the same means as 'primitive' folk psychology. On this account, there is no problem in explaining our willingness to attribute irrational beliefs, despite the fact that there are grounds for viewing mental explanation as having a normative basis, in virtue of the impact of *the auxiliary law* and *the coherence principle.*

A question remains to be addressed. Why does Dennett not endorse a variation of the strategy advocated here as a solution to the problem of how to

accommodate irrational belief amongst a normatively-driven folk psychology, and in turn, Intentional Systems Theory?

Dennett explicitly derides the sort of approach advocated here on the grounds that permitting exceptions from the ideal of perfect rationality is like having a rule in chess which states that we are permitted to break the ordinary rules of chess a predetermined number of times per game.(Dennett 1980) Despite this cryptic remark, it seems that Dennett has misgivings with the strategy because it edges ever so slightly in a direction which is at odds with his inviolable assumption that psychological explanation is necessarily instrumentalistic in virtue of the thoroughly normative basis of intentional state attribution.

It is obviously false that the bases of mental explanation are thoroughly normative. While the function of the coherence principle is to ensure the projectability of the sets of intentional states we attribute. The principle also provides a basis for understanding the way in which the normative bases of interpretation work against the empirical bases of interpretation to the betterment of interpretations. I have already argued that mental explanation is underpinned by an implicit understanding of the normative relations between state types. Despite the presumption of upholding these relations in the course of interpretation, mental explanation permits the integration of empirical data as a basis for recognising the types of cases where one should make an exception to the core laws. Thus, while the presumption of maximising the coherence of the sets of intentional states we attribute prevents the attributions of sets of intentional states which will not support predictions of behaviour, knowledge of the sorts of conditions under which the core laws are prone to exception constrains the application of the coherence principle.

The suppression the principle allows the application of knowledge about the regularity of patterns of intentionally and non-intentionally described behaviour to have an impact upon the interpretive process.

My conjecture is that Dennett (at least, at certain moments) supposes that opening up the possibility of the rationalisation of the behaviour of individuals by appeal to cognitive short-comings puts in jeopardy his presumption that the attribution of intentional states rests on no assumptions about the physical/internal mechanisms which realise the bases of behaviour. Despite this concern, it is clear that the attribution of any cognitive habits, or even of intentional states, presupposes the existence of models which actual physical brain structures may correspond to in some manner. So it is unclear why instrumentalism with regard to cognitive short-comings is unacceptable. Indeed, such instrumentalism is consistent with Dennett's claim that "Intentional Systems Theory just deals with the performance specifications of believers while remaining silent on how the systems are to be implemented."(Dennett 1987, p.59) Thus, it appears that Dennett's prejudice against the sort of response to Stich that I have proposed is unwarranted, and possibly symptomatic of deeper problems with the proposal that the normative character of the folk theory provides some sort of insurance against elimination.

So far I take myself to have shown, first, that empirical evidence supports the claim that our capacity for intentional interpretation is knowledge-based, and, second, that this knowledge base possesses the features generally supposed to be central to the correct conception of a scientific theory. A further result derived from this account of intentional psychology is that the sort of the theory which the folk theory appears to be suggests why it is that we can attribute irrational beliefs despite

the theory's normative basis. The question of whether the normative basis of mental explanation implies the ineliminability and/or irreducibility of intentional psychology will be set aside until chapter 4.

At this point I wish to continue my investigation of the nature of interpretation by schematising parts of a formal foundation for a method for interpreting intentional behaviour. This method, which differs somewhat from a method which would emulate *the machine model of deduction*, incorporates the assumption that an agent's set of intentional states must possess an agent relative degree of coherence.

Chapter 2:

How to Measure the Coherence of a Set of Intentional States.

In the preceding chapter, I argued that we should suppose that individuals possess varying capacities to detect inconsistencies among their belief sets, and corresponding capacities to act in accordance with the classical consequences of their beliefs. The presupposition that motivates this claim is that we ordinarily do not wish to attribute to any individual a self-conscious apprehension of any particular inconsistency among their belief set. In chapter one, I also suggested a sense in which we may regard the basis for intentional state attribution as normative, and claimed that we should adopt as a principle of the methodology of intentional psychology the aim of maximising the coherence of the sets of intentional states we attribute. Evidently there will be a correspondence between the ability we attribute to an individual to detect inconsistency among his belief set, and the degree of our adherence to the *coherence principle*.

I now wish to embark on a detailed exploration of how we might provide formal definitions of measures of coherence and classical closure which reflect the conclusions reached in the preceding chapter. The objective is to show how such precisely defined measures might factor in a method of interpretation which maintains allegiance to the core laws of intentional psychology inasmuch as the application of such measures respect the core laws by providing a scheme for the rationalising instances of their exception. The idea is that we will tolerate the attribution of particular degrees of inconsistency on the assumption that there are degrees of consistency which would lie beyond an agent's ability to detect. The

hope is also to give sense to the demand that we maximise the consistency of the sets of intentional states we attribute to others, by providing a measure by which we can grade the coherence of a set of intentional states.

Despite the advertised topic of this chapter, "*how* to measure the coherence of a set of intentional states", I will also address the question as to *why* we ought to measure the coherence of sets of intentional states for the sake of restricting their incoherence. With regard to this question I will draw a distinction between *methodological* and *strategic* considerations, which counsel us to attribute rationality to the individuals we interpret. Strategic considerations come in for discussion in section 2.3; throughout section 2.1, I will attempt to discover a measure of coherence that accords with the methodological grounds for maximising the coherence of the sets of intentional states we attribute. Throughout section 2.2, I will describe several interesting connexions between the measures of coherence introduced in section 2.1.

§2.0 On the Application of Coherence Measures in the Process of Interpretation: Preliminary Remarks.

The attribution of a coherence measure as a constraint on the sets of intentional states we attribute will be agent-relative and will function on two levels:

(1) Such measures will function at the time when we assign a set of *basic behaviour-explaining* intentional states. A set of *basic behaviour-explaining* intentional states (or a set of *basic explanatory intentional states*) is simply a set of intentional states which are sufficient for explaining the behaviour of a subject up to

the present.[1] On this level, acceptable sets of attributions are constrained by a coherence requirement, in the form of an agent relative *ceiling* (or *floor*, depending on the measure) on the incoherence of the sets of intentional states we may attribute. Moreover, the role of the attributed coherence restriction is obviously negative, since the restriction merely rules out the attribution of some sets of intentional states.

(2) As consequences of our basic attributions, various intentional states can be thought of as implicitly present in the mind of the subject of interpretation. These implicit intentional states are attributed according to a closure condition that preserves the coherence of the agent's basic explanatory intentional states. Such implicit beliefs establish possibilities for the prediction of the behaviour of the subject which do not require the augmentation of the agent's basic explanatory intentional states.

The second level at which the attribution of a coherence restriction functions invites the development of paraconsistent logics. I shall return to this problem in chapter 3. The remainder of this chapter concerns the attribution of basic explanatory intentional states.

---

[1] In the ideal and under optimal conditions, such a set would explain *all* of the individual's behaviour. In the ideal but under less optimal conditions, the set would explain all of the subject's observed behaviour, and would implicitly incorporate assumptions about the probable history of the individual. For example, certain intentional states would be attributed to explain the individual's dress, and his knowing how to speak etc. In practice, it is rarely worthwhile to attribute a full-blown set of basic explanatory intentional states. It is obvious that in typical encounters such as passing an innocuous pedestrian upon the sidewalk we devote little energy to attributing basic explanatory intentional states. We are satisfied if we have determined that the active intentional states of the subject are not such as to lead him into the trajectory of our own path.

§2.1 Coherence Maximisation as a Dictate of the Methodology of Interpretation:

With regard to the question, "by what measure ought we to maximise the coherence of the sets of intentional states we attribute to others?", it will shortly become evident to the reader that there are many ways to give sense to the demand. In accordance with this observation, the attempt will be made to show that some measures are more suitable than others. What follows can be thought of as providing a menu of options whereby we can measure the coherence of a set of intentional states. I will also try to elucidate the nature of these measures by revealing some of the properties characteristic of sets which possess them. The precept of the investigation will be that relatively to an agent, a set of intentional states is coherent to the degree that it resists the derivation of contradictions.

If one grants, for the sake of argument, what was claimed in section 1.3.3 concerning why one ought to temper our attributions of inconsistency, then one might suppose that we ought to define the coherence of a belief set in accordance with its resistance to the derivability of contradictions in face of something like a theorem-proving algorithm. While the machine model appears adequate, it lacks elegance and appears to have limitations as a basis for the development of a logic of belief attribution (the project postponed until chapter 3). For this reason, I shall survey several other proposals for the measurement of coherence. I begin by considering a measure which has received attention by paraconsistentist logicians.(by Schotch and Jennings in (1980) and (1989), by Jennings and Schotch

in (1984), and by Apostoli and Brown in (1995)) The plan is to evaluate the suitableness of this measure for the present application.

§2.1.1 Level of Incoherence[2]:

Informally, the incoherence level of a set of sentences is equal to the size of the least, covering family of consistent subsets. This definition can also be given using the notion of an n-partition. The notion of an n-partition (and several derivative notions) will also prove useful in expressing other concepts in chapter 3.

Definition 1: An *n-partition*, π, of a set, Σ, is a set of sets

$$\{a_1, ..., a_n\} \text{ (where } n \geq 1): \forall i: a_i \neq \varnothing \ \& \ (\forall i, j: \ i \neq j \Rightarrow a_i \cap a_j = \varnothing) \ \& \ \cup\pi = \Sigma.$$

Definition 2: $\prod_n(\Sigma)$ is the set of n-partitions of Σ. That is:

$$\prod_n(\Sigma) = \{ \pi = \{a_1,...,a_n\} \mid \forall i, j: i \neq j \Rightarrow a_i \cap a_j = \varnothing \ \& \ \cup\pi = \Sigma \}.$$

Definition 3: $\prod_{n*}(\Sigma)$ is the set of partitions of Σ which have n or more members.

Definition 4: $\prod(\Sigma)$ is the set of all partitions of Σ.

Using the notion of an n-partition, Jennings & Schotch's measure is defined as follows:

Definition 5: *The incoherence level of a set* Σ, called 'level of Σ' for short and written $\ell(\Sigma)$, is equal to the size of the least partition of Σ into subsets none of which

---

[2] This measure was introduced by P.K. Schotch and R.E. Jennings.

are inconsistent. (If there is no such partition, the level of a set is some arbitrarily large value, ∞.)

Formally, $\ell(\Sigma) = \min\{ n \mid \exists A \in \prod_n(\Sigma): \forall a \in A: a \not\vdash_{PL} \bot \}$, if $\bot \notin \Sigma$, and
$= \infty$ else.

Sets of level m can be generated by the following schema:

Schema 1:

$$\Gamma_1 = \{ p_1 \}.$$

and generally $\Gamma_i = \Gamma_{i-1} \cup \{ p_i \wedge \neg p_1 \wedge \ldots \wedge \neg p_{i-1} \}$.

For example:

$$\Gamma_2 = \{ p_1, p_2 \wedge \neg p_1 \}.$$

$$\Gamma_3 = \{ p_1, p_2 \wedge \neg p_1, p_3 \wedge \neg p_1 \wedge \neg p_2 \}.$$

Theorem 1: $\forall m > 0: \ell(\Gamma_m) = m$.

The idea of measuring inconsistency by incoherence level is consonant with Davidson's prescribed method for rationalising the attribution of inconsistency. Davidson claims that in such cases we must attribute *cognitive partitions* between the inconsistent intentional states. Certainly some version of this principle should be conceded. Thus, according to the Davidsonian view, minimising the level of a set of intentional states minimises the amount of work we must do in rationalising the irrationality of an agent. The level of a set of intentional states minus one is equal to the number of cognitive partitions that must be attributed.

Despite the consonance of level of incoherence with Davidson's recommendations, the measure does not distinguish sets such as a and b.

$$a = \{ p, \neg p \}$$

$$b = \{ p, p \supset q, q \supset r , \neg r \}$$

Intuitively, it would be easier to rationalise the attribution of b than a. One would like to reflect this intuition in a formally crisp way.

§2.1.2 Corruption:

A measure of incoherence that provides for the differentiation of a and b is what I will call the corruption of a set. Informally, a set is thought to be corrupt if a large proportion of its subsets are inconsistent. More formally, given a set, Σ, we may refer to the fraction of its subsets that are inconsistent as the *corruption* of Σ, $C(\Sigma)$. That is:

Definition 6: $C(\Sigma) = |\{ c \subseteq \Sigma \mid c \vdash_{PL} \perp \}| / |\wp(\Sigma)|$.[3]

Some Examples:

let $a = \{p, \neg p\}$

$\wp(a) = \{ \varnothing, \{ p\}, \{\neg p\}, \mathbf{\{ p, \neg p\}} \}$.[4]

$C(a) = 1/4$.

let $b = \{ p, p \supset q, q \supset r , \neg r \}$

---

[3] $|\Sigma|$ is the number of elements of Σ. $\wp(\Sigma)$ is the *powerset* of Σ, that is, the set of all subsets of Σ. For ease of reference, these definitions, along with some others, are repeated in appendix A.

[4] Inconsistent sets are in bold.

$\wp(b) = \{\varnothing, \{p\}, \{p \supset q\}, \{q \supset r\}, \{\neg r\}, \{p, p \supset q\}, \{p, q \supset r\}, \{p, \neg r\},$
$\{p \supset q, q \supset r\}, \{p \supset q, \neg r\}, \{q \supset r, \neg r\}, \{p, p \supset q, q \supset r\}, \{p, q \supset r, \neg r\},$
$\{p, p \supset q, \neg r\}, \{p \supset q, q \supset r, \neg r\}, \mathbf{\{p, p \supset q, q \supset r, \neg r\}}\}$

$C(b) = 1/16.$

Besides reflecting a relevant difference between a and b, this measure seems in some respects to be in accord with the demands that prompted our search. The corruption of a set is a measure of the ease with which we may draw contradictions from it. Moreover, the presumption of our search was that only consistent subsets of a set of intentional states can be considered candidates for active determinants of their possessor's behaviour. In turn, the complement of a set's corruption (that is, $1 - C(\Sigma)$) is the proportion of the set's subsets eligible to be active determinants of an agent's behaviour.

While the measure, *corruption*, is in accord with the dictate that no inconsistent subset of a set of intentional states can be an active determinant of behaviour, there are other measures of coherence which differentiate a and b. One such measure is also in accord with plausible intuitions about the systematicity of an agent's capacity to detect inconsistency among his intentional states.

It seems incorrect to suppose that an agent's behaviour may be actively determined by any consistent subset of his set of intentional states. Rather it seems that an agent can act on the basis of any subset of his intentional states that he can synchronically grasp the sense of. Since we ought to suppose that agents cannot grasp the sense of any inconsistent subset of their belief set without thereby seeing that at least one element of the set ought to be repudiated, there must be some

property (such as a degree of complexity) that prevents an agent from synchronically grasping the sense of the inconsistent sets of beliefs which he persists on holding. Similarly, it seems that an agent would also be unable to synchronically grasp the sense of any consistent subset of his belief set possessing the same property (that degree of complexity) as the inconsistent subsets of his belief set.

§2.1.3 Dilution of Incoherence:

Definition 7: *The dilution of incoherence of* Σ, called 'dilution of Σ' for short and written $\mathcal{d}(\Sigma)$, is the size of the smallest inconsistent subset of Σ. (If the set is consistent, its dilution is some arbitrarily large value, ∞.)[5]

$$\text{Formally, } \mathcal{d}(\Sigma) = \min\{ n \mid \exists a \subseteq \Sigma : |a| = n \ \& \ a \vdash_{PL} \bot \}, \text{ if } \Sigma \vdash_{PL} \bot, \text{ and}$$
$$= \infty \text{ else.}$$

Recall the sets $a = \{ p, \neg p \}$, and $b = \{ p, p \supset q, q \supset r , \neg r \}$. $\mathcal{d}(a) = 2$ and $\mathcal{d}(b) = 4$.

Sets of dilution n (where $n > 1$), can be generated by the following schema:

Schema 2:

$$\Delta_n = \{ p_1, p_1 \supset p_2, \ldots , p_{n-2} \supset p_{n-1}, \neg p_{n-1} \}.$$

Theorem 2: $\forall n > 1: \mathcal{d}(\Delta_n) = n$.

---

[5] To my knowledge, this definition was first given in print in an unpublished work by R.E. Jennings, called *Leibnizian Semantics* (1984), though a similar, but not identical definition, was given by Henry Kyburg in his paper *Conjunctivitis* (1970).

Since more dilute inconsistencies are more difficult to detect, it is evident that the measure, *dilution of incoherence*, can be used to discriminate sets in proportion to how realistic it is to suppose that an individual would fail to 'notice' their inconsistencies. Thus, the use of dilution as a restriction on the sets of intentional states we would attribute to an agent would be suitable as a measure of the acuity and vigilance of the agent with regard to the maintenance of his intentional states. But the aptness of this measure for the proposed application transcends this fact.[6]

The measure's promise for the proposed application lies in what it implicitly measures: a set's threshold (of size) below which all subsets of the set are consistent.

Theorem 3: $\forall \Sigma,n\colon\ \mathcal{d}(\Sigma) = n \Rightarrow (\forall a \subseteq \Sigma\colon |a| < n \Rightarrow a \nvdash_{PL} \bot)$.

A more general version of this theorem can also be given.

Theorem 4: $\forall \Sigma,n\colon\ \mathcal{d}(\Sigma) \geq n \Leftrightarrow (\forall a \subseteq \Sigma\colon |a| < n \Rightarrow a \nvdash_{PL} \bot)$.

The point of theorem 3 is simple but of great consequence. Given the informal measure of coherence, according to which the coherence of a set is a function of its resistance to the derivation of contradictions, it is evident that dilution is one mode of such resistance. All subsets of a set smaller than the set's dilution are *safe subsets*.

---

[6] Another consequence of using dilution as the measure for the proposed application is that it would preserve a Davidsonian conviction that despite the partitioning of the mind required to explain irrationality we must suppose that inconsistent beliefs, thus partitioned, must still belong to "strongly overlapping territories", since the inconsistent content of the intentional states is still constituted by the relations in which the states stand to other intentional states.(Davidson 1986, p.91-2) The definition of dilution satisfies Davidson's demands since a single sentence which is a member of an n-diluted set need be conceived as separated from the constituent members of the set taken as (n -1)-tuples.

We can take their consequences without encountering any contradictions. Thus, a significant effect of adopting dilution of incoherence for the proposed application would be that we thereby ensure that the sets of intentional states we attribute have the sort of integrity which makes such sets sound as bases for making predictions of behaviour of their possessors. By placing a higher floor on dilution we ensure that we may deem larger subsets of an individual's set of intentional states to be *active* determinants of behaviour. A relationship between the dilution of sets and their corruption can also be established.

Theorem 5: $\forall\Sigma,n:\ \mathscr{d}(\Sigma) \geq n \Rightarrow C(\Sigma) \leq |\{c \subseteq \Sigma \mid |c| \geq n\}| \ / \ |\wp(\Sigma)|$.[7]

In this theorem the set $\{c \subseteq \Sigma \mid |c| \geq n\}$ represents the set of subsets of a set whose size is greater than that for which we have a guarantee of consistency.

And, of course, $|\{c \subseteq \Sigma \mid |c| \geq n\}| = \sum_{i=n}^{|\Sigma|} \binom{|\Sigma|}{i}$.

An illustration:

let $\Sigma = \{\alpha, \beta, \gamma, \delta\}$, and $\mathscr{d}(\Sigma) \geq 3$.

Then $\wp(\Sigma) = \{\varnothing, \{\alpha\}, \{\beta\}, \{\gamma\}, \{\delta\}, \{\alpha, \beta\}\ \{\alpha, \gamma\}$
$\{\alpha, \delta\}, \{\beta, \gamma\}, \{\beta, \delta\}, \{\gamma, \delta\}, \mathbf{\{\alpha, \beta, \gamma\}, \{\alpha, \beta, \delta\}}$,
$\mathbf{\{\alpha, \gamma, \delta\}, \{\beta, \gamma, \delta\}, \{\alpha, \beta, \gamma, \delta\}}\}$[8]

And $C(\Sigma) \leq 5/16$.

---

[7] It can also be shown that knowing the corruption of a set, Σ, allows us to compute a lower limit of its dilution: **Theorem:** $\forall\Sigma: C(\Sigma) \leq m \Rightarrow \mathscr{d}(\Sigma) \geq \left\lfloor \log_2 (1/m) \right\rfloor$. The proof is in appendix B.

[8] Subsets which could be inconsistent are in bold.

§2.1.5 More On the Aptness of the Measure, *Dilution of Incoherence*:

In chapter 1, I made the distinction between *active* and *inactive* intentional states, and argued that a distinction such as this must be embraced so long as we wish to attribute inconsistent sets of intentional states. For the purposes of interpreting (and, in particular, predicting) intentional behaviour, it is recommended that we also distinguish between the sets of intentional states of an agent that have the *potential* of being active, and those which do not. In the process of interpretation, and the prediction of intentional behaviour, we select the determinants of an agent's behaviour from among an agent's set of *potentially active sets of intentional states.* I suggested that we suppose that the most relevant set of eligible states generally become active according to varied circumstances.

On the supposition that we cannot regard any inconsistent subset of an agent's set of intentional states as potentially active, one might suppose that we can regard any subset of a set of intentional states, Σ, of size $\mathcal{A}(\Sigma)$-1 as potentially active. In fact, this assumption may be too liberal. In addition to the imperative that we not attribute a potentially active set of intentional states which is inconsistent, we should avoid the attribution of pairs of potentially active sets of intentional states whose union is inconsistent. If we do attribute such pairs we are bound to find ourselves in situations where the set of potentially active sets of intentional states we have attributed is inadequate as a basis for making predictions of the behaviour of the agent. This will arise in virtue of the fact that *for any* pair of the potentially active sets of intentional states whose union is inconsistent, there are bound to be situations in which each of the potentially active sets are equally relevant to the

agent's situation. Thus, we will be unable to predict which set is likely to become active. In addition to this, because the union of the two sets is inconsistent it may well be that each of the two sets would dictate courses of action which are opposed and/or divergent to the other.

The following theorem will tell us, given the dilution of an agent's set of intentional states and the dilution we wish to maintain for the union of the agent's set of potentially active sets of intentional states, an upper limit on the size of elements of the agent's set of potentially active intentional states. That is, if Σ is an agent's set of intentional states, and if $\mathcal{d}(\Sigma) = r$, and we wish to allow the systematic aggregation of elements of Σ to produce an extension of Σ, Σ^*, such that the $\mathcal{d}(\Sigma^*) \geq s$, then we may generate Σ^* according to the rule, *n-ary aggregation*, as follows:

(The value of n can be determined via theorem 6.)

[n-AG] $$\frac{\alpha_1 \in \Sigma \ \& \ \ldots \ \& \ \alpha_n \in \Sigma}{\alpha_1 \wedge \ldots \wedge \alpha_n \in \Sigma^*}$$

Theorem 6: $\forall\Sigma$: $\forall r > 1$: $\mathcal{d}(\Sigma) = r$ & $n < r/(s-1)$ &

$$\Sigma^* = \{ \Sigma^* \mid (\alpha_1 \in \Sigma \ \& \ \ldots \ \& \ \alpha_n \in \Sigma) \Rightarrow (\alpha_1 \wedge \ldots \wedge \alpha_n \in \Sigma^*) \} \Rightarrow \mathcal{d}(\Sigma^*) \geq s.$$[9]

[9] A similar rule of aggregation will hold for sets of statements which have been assigned measures of probability. Assuming the independence of the probabilities assigned to elements of a set, we may aggregate as above. If $\forall\alpha \in \Sigma$, PROB(α) $\geq p$, and if we wish to accept all statements of a probability greater or equal to a, then $n = \left\lfloor \log_p a \right\rfloor$.

Thus, for example: if $\mathcal{d}(\Sigma) = 3$ and $n = 2$, then $\mathcal{d}(\Sigma^*) = 2$, and

if $\mathcal{d}(\Sigma) = 10$ and $n = 2$, then $\mathcal{d}(\Sigma^*) = 3$.[10]

Proof of Theorem 6:

Assume not, That is:

$$\exists\Sigma: \mathcal{d}(\Sigma) = r \;\&\; n < r/(s-1) \;\&$$
$$\Sigma^* = \{\, \Sigma^* \mid (\alpha_1 \in \Sigma \;\&\; \ldots \;\&\; \alpha_n \in \Sigma) \Rightarrow (\alpha_1 \wedge \ldots \wedge \alpha_n \in \Sigma^*) \,\} \;\&\; \mathcal{d}(\Sigma^*) < s.$$

Since $\mathcal{d}(\Sigma^*) < s$, $\exists b \subseteq \Sigma^*: |b| < s \;\&\; \cup b \vdash_{PL} \perp$.

Moreover, $\exists b \subseteq \Sigma^*: |b| \leq (s-1) \;\&\; \cup b \vdash_{PL} \perp$.

We note that each element of Σ^* is composed of n conjoined elements of Σ.

Thus, $\exists b \subseteq \Sigma: |b| \leq (s-1)(n) \;\&\; \cup b \vdash_{PL} \perp$.

But since $n < r/(s-1)$, $r > (s-1)(n)$.

Therefore, $\exists b \subseteq \Sigma: |b| < r \;\&\; \cup b \vdash_{PL} \perp$.

But $\mathcal{d}(\Sigma) = r$. Therefore, $\forall b \subseteq \Sigma: |b| < r \Rightarrow \cup b \nvdash_{PL} \perp$.

It follows that $\exists b \subseteq \Sigma: |b| < r \;\&\; \cup b \vdash_{PL} \perp \;\&\; \cup b \nvdash_{PL} \perp$, which is absurd.

Despite the generality of theorem 6, the particularity of the following corollary is of the most use, since problems of the unprojectibility of the behaviour of an agent owing to the inconsistency of his intentional states does not seem to arise so long as we attribute sets of potentially active sets of states which have a dilution of at least three.

---

[10] In Theorem 6 the clause, $n < r/(s-1)$, could be replaced by the equivalent, $n \leq \left\lceil r/_{s-1} \right\rceil - 1$.

Corollary 1: $\forall\Sigma: \forall r > 1: \mathcal{d}(\Sigma) = r \;\&\; n = \lceil r/2 \rceil - 1 \;\&$

$$\Sigma^* = \{\, \Sigma^* \mid (\alpha_1 \in \Sigma \;\&\; \ldots \;\&\; \alpha_n \in \Sigma) \Rightarrow (\alpha_1 \wedge \ldots \wedge \alpha_n \in \Sigma^*) \,\} \;\Rightarrow \mathcal{d}(\Sigma^*) \geq 3.$$[11]

This corollary upholds the justification for the claim that there is a relationship between a set's dilution and the upper limit on the desirable size of elements of an agent's set of potentially active sets of intentional states. It follows that the dilution of an agent's set of intentional states is apt as a measure of the integrity of the set as a basis for predicting the agent's behaviour. Indeed, a higher dilution allows us to deem larger subsets of an agent's set of intentional states to be potentially active. In turn, allowing larger subsets of an agent's set of intentional states to be potentially active entails that our interpretation will support a greater number of principled predictions about how the agent will behave in various circumstances. These points can be illustrated using simple examples.

Suppose we place two agents in a situation where each agent must perform one of three actions, X,Y, or Z. Suppose further that the first agent, A_1, has a set of preferences of dilution two. We let A_1's preferences be { X > Y, Y > X }. That is, A_1 would prefer to X than to Y, and also prefers to Y than to X. On the other hand, suppose that a second agent, A_2, has a set of preferences of dilution 3. Let A_2's preferences be { X > Y, Y > Z, Z > X }. Finally, assume that elements of A_1's and A_2's sets of potentially active sets intentional states are all 1-membered. It is clear that in the situation in which each agent may perform one of three actions, X,Y, or Z, there is no principled way of making predictions about which action either agent will

---

[11] $\lfloor x \rfloor$ is equal to x rounded down to the nearest integer. $\lceil x \rceil$ is equal to x rounded up to the nearest integer.

perform. However, suppose that instead of putting the two agents in a situation where they may perform one of three of X, Y, and Z, we place them in situations where they may choose to perform one of two actions. Now it seems that there is a difference in our ability to make principled predictions of the behaviour of A_1 and A_2. Assuming that the most relevant potentially active intentional state becomes active, we can predict A_2's behaviour in all of the three possible cases, yet we still cannot make principled predictions of A_1's behaviour. Moreover, the fact that we can now attempt to predict A_2's behaviour seems to be a function of the dilution of his set of potentially active preferences. Consider what happens if we decrease the dilution of the set of his potentially active preferences to two, by increasing the size of elements of his set potentially active sets of preferences to two. A_2 now has the following set of potentially active preferences: { { X > Y, Y > Z }, { X > Y, Z > X }, { Y > Z, Z > X }}. We can no longer make principled predictions of A_2's behaviour. Indeed, for each situation in which A_2 is placed, he possesses a pair of potentially active preferences counseling him to make opposite choices. Let us now return to the case where an agent has the potential of performing one of three actions, X, Y, or Z. Suppose that we place an agent, A_3, in such a situation, and suppose that she possesses a set of preferences of dilution five: { X > Y, Y > Z, Z > S, S > T, T > X }. Now it must be observed that in order to decide between three options, two statements of preference must be active. Thus, let us suppose that the size of elements of A_3's sets of potentially active preferences are 2-membered. The set of A_3's potentially active preferences is then: { { X > Y, Y > Z }, { X > Y, Z > S }, { X > Y, S > T }, { X > Y, T > X }, { Y > Z, Z > S }, { Y > Z, S > T }, { Y > Z, T > X },

{ Z > S, S > T }, { Z > S, T > X }, { S > T, T > X } }. It is clear that the first member of this set would compel A_3 to choose to perform X. Moreover, no other element of the set contradicts this choice. Finally, I should comment on the fact that none of A_3's potentially active sets of preferences are inadequate as a basis for making a prediction about what act she should perform in a situation where she may perform one of X, Y, or S. What we would like to do to remedy the situation is to increase the size of the elements of A_3's set of potentially active preferences, thereby allowing one of these sets to have enough content to compel an action. However, increasing the size of the elements of A_3's potentially active preferences also has the effect of decreasing the dilution of these sets from three to two, having the effect that there would now be pairs of elements of A_3's set of potentially active preferences which compel contradictory actions. The final point is that this problem would not arise if we reinterpreted A_3's preferences by removing T > X from the set and adding T > U and U >X, thereby making the dilution of her set of preferences seven, rather then five. We could then (with impunity) increase the size of elements of A_3's set of potentially active preferences to three which would allow us to predict that she will choose to perform X, in the situation where she has the option of performing X, Y, or S. Thus, the general moral is that maximising the dilution of the sets of intentional states we attribute, maximises the integrity of these sets as a basis for predicting the behaviour of their possessors. Of course, this line of thought deserves greater attention. Nevertheless, perhaps, the general point has been sufficiently made.

§2.2 Level and Dilution Compared:

Setting aside our interest in the application of the measure, *dilution*, it is of interest to notice that the definition of dilution captures a formal notion of coherence that nicely complements the notion of level of incoherence. Such a comparison also makes clear the basis of the relevant difference between dilution and level which makes dilution suitable for the present application and level not. The force of the contrast between the two notions is put in clearer terms if we recognise that a set's level of incoherence is simply a function of the compounding of inconsistency via the conjunction of distinct truth-functional atoms (or formulae that behave as such). The following schema sufficiently illustrates this fact.

Schema 3:

Where $m = \lceil \log_2 n \rceil$,

and where $*_x p_y = \neg p_y$ if the y^{th} digit from the right of Bin(x - 1) = 1, and [12]
$= p_y$ else.

$$\Gamma^*_n = \{ *_1 p_1 \wedge \ldots \wedge *_1 p_m, \ldots, *_n p_1 \wedge \ldots \wedge *_n p_m \}.$$

Some instances of schema 3:

$$\Gamma^*_2 = \{ p_1 \wedge p_1, \neg p_1 \wedge \neg p_1 \}, \qquad \Gamma^*_3 = \{ p_1 \wedge p_2, \ p_1 \wedge \neg p_2, \neg p_1 \wedge \ p_2 \},$$

$$\Gamma^*_5 = \{ \ p_1 \wedge p_2 \wedge p_3, \ \ p_1 \wedge p_2 \wedge \neg p_3, \ \ p_1 \wedge \neg \ p_2 \wedge p_3, \ \ p_1 \wedge \neg \ p_2 \wedge \neg \ p_3, \ \ \neg p_1 \wedge p_2 \wedge p_3 \}.$$

Theorem 7: $\forall n > 0: \ell(\Gamma^*_n) = n$.

---

[12] Bin(x) is the binary representation of a natural number x.

Distinct from a set's level, a set's dilution is also a function of the compounding of distinct truth-functional atoms (or again formulae behaving as such), where the assertive strength of each sentence is weakened by the disjunction of such formulae, rather than by their conjunction. The following schema illustrates the point:

Schema 4:

Where $m = \lceil \log_2 n \rceil$,

and where $*_x p_y = \neg p_y$ if the y^{th} digit from the right of Bin(x - 1) = 1, and
$= p_y$ else.

and where $[,_x] =$, if $x < n$, and
$= \wedge$ else.

$$\Delta^*_n = \{ (*_1 p_1 \vee \ldots \vee *_1 p_m)[,_1] \ldots [,_{2^m-1}](*_{2^m} p_1 \vee \ldots \vee *_{2^m} p_m) \}.$$

Some instances of schema 4:

$$\Delta^*_2 = \{ p_1 \vee p_1, \neg p_1 \vee \neg p_1 \}, \quad \Delta^*_3 = \{ p_1 \vee p_2, p_1 \vee \neg p_2, (\neg p_1 \vee p_2) \wedge (\neg p_1 \vee \neg p_2) \},$$

$$\Delta^*_5 = \{ p_1 \vee p_2 \vee p_3, \ p_1 \vee p_2 \vee \neg p_3, \ p_1 \vee \neg p_2 \vee p_3, \ p_1 \vee \neg p_2 \vee \neg p_3,$$
$$(\neg p_1 \vee p_2 \vee p_3) \wedge (\neg p_1 \vee p_2 \vee \neg p_3) \wedge (\neg p_1 \vee \neg p_2 \vee p_3) \wedge (\neg p_1 \vee \neg p_2 \vee \neg p_3) \ \}.$$

Theorem 8: $\forall n > 1: \mathcal{d}(\Delta^*_n) = n$.

While schema 4 is somewhat cumbersome, it is interesting inasmuch as the duals of its instances (that is, where all instances of $\wedge$ are replaced by $\vee$, and all instances of $\vee$ are replaced by $\wedge$) are of level n. Although sets do not generally possess the dilution and level, respectively, equivalent to the level and dilution of their dual, the

equivalence in the preceding schema illustrates the relation between dilution and disjunction on the one hand, and level and conjunction on the other. The contrast of the two measures is also represented by a pair of theorems, which reflect the number of distinct truth-functional atoms required to construct sets of respective magnitudes of level and dilution.

Theorem 9: $\forall\Sigma\colon \ell(\Sigma) = n \Rightarrow \mathit{at}(\Sigma) \geq \lceil \log_2 n \rceil$.[13]

Theorem 10: $\forall\Sigma\colon d(\Sigma) = n \Rightarrow \mathit{at}(\Sigma) \geq \lceil \log_2 n \rceil$.

Proof of theorem 9: The result is straightforward. As is well known, the set of equivalence classes of propositional models for a universe of n truth-functional atoms is 2^n. Moreover, the number of consistent but pairwise inconsistent sentences (or sets of sentences) that could be constructed using n atoms is also 2^n. Each sentence would be valid in exactly one of the equivalence classes of models for the set's atoms, and no pair of sentences would be valid in any of these equivalence classes.

Proof of theorem 10: We need only recognise that the dilution of a set, Σ, is equal to the size of the least subset, s, of Σ that is not satisfied by any propositional model. The maximum size of s would be equal to the set of equivalence classes of propositional models for its truth-functional atoms, since the largest inconsistent set that possesses no inconsistent subsets will be composed of sentences that are each valid in all but one element of the set of equivalence classes of models for its atoms. Thus, where the number of atoms in s is n, the number of equivalence classes and the maximum size of s is 2^n.

---

[13] $\mathit{at}(\Sigma)$ is the number of distinct truth-functional atoms which appear in Σ.

§2.2.1 The Independence of Dilution and Level:

The schema which follows illustrates the independence of particular dilutions and levels of incoherence. The independence claim is demonstrated by the sketch of a procedure for generating sets of any dilution and level of incoherence, save for two special cases:

Theorem 11: $\forall\Sigma$: $\ell(\Sigma) = \infty \Leftrightarrow \mathcal{d}(\Sigma) = 1$.

Theorem 12: $\forall\Sigma$: $\ell(\Sigma) = 1 \Leftrightarrow \mathcal{d}(\Sigma) = \infty$.

The respective sets, $\Sigma_{m|n}$, which could be generated by a procedure of the following kind, will be of level m, and of dilution n.

That is, $\forall m > 1$: $\forall n > 1$: $\ell(\Sigma_{m|n}) = m$ & $\mathcal{d}(\Sigma_{m|n}) = n$.

Schema 5:

The procedure functions by listing the names of the sentences which will compose $\Sigma_{m|n}$, and then describing the sentences.

For all m and n greater than one, we stipulate the size of $\Sigma_{m|n}$:

$$|\Sigma_{m|n}| = (m - 1)(n - 1) + 1.$$

The elements of $\Sigma_{m|n}$ are named $\alpha_{1_{m|n}}, \ldots, \alpha_{(m-1)(n-1)+1_{m|n}}$.

The sentences composing $\Sigma_{m|n}$ are subsequently described as conjunctions generated with reference to a schema for generating n-membered sets of dilution n. We can use schema 4.

We take the first $\binom{(m-1)(n-1)+1}{n}$ elements of an ordered set of *variations* of the respective instances of schema 4. All variations of each instance are assumed to be pairwise disjoint with respect to their truth-functional atoms. The elements of each variation are also assumed to be ordered.

Next we generate a set of sets $\sum^{*m|n} = < a_{1_{m|n}}, \ldots, a_{(m-1)(n-1)+1_{m|n}} >$ according to a procedure which assigns a respective variation of schema 4 to each element of an ordering of the n-tuples of $\sum^{*m|n}$. Moreover, the ith element of the respective ordered variation of schema 4 is assigned to the ith element of the ordered n-tuple.

Finally, the set $\sum_{m|n} = < \alpha_{1_{m|n}}, \ldots, \alpha_{(m-1)(n-1)+1_{m|n}} >$ is generated by a procedure which assigns to each member of $\sum_{m|n}$, a conjunction which corresponds to the elements of the corresponding element of $\sum^{*m|n}$. (For example, if $a_{i_{m|n}} = \{ p_1 \vee p_2, p_3 \vee p_4 \}$, then $\alpha_{i_{m|n}} = (p_1 \vee p_2) \wedge (p_3 \vee p_4)$.)

An example:

$\sum_{3|3} = \{ \alpha_{1_{3|4}}, \alpha_{2_{3|4}}, \alpha_{3_{3|4}}, \alpha_{4_{3|4}}, \alpha_{5_{3|4}} \}$.

$\alpha_{1_{3|4}} = (p_1 \vee p_2) \wedge (p_3 \vee p_4) \wedge (p_5 \vee p_6) \wedge (p_7 \vee p_8) \wedge (p_9 \vee p_{10}) \wedge (p_{11} \vee p_{12})$

$\alpha_{2_{3|4}} = (p_1 \vee \neg p_2) \wedge (p_3 \vee \neg p_4) \wedge (p_5 \vee \neg p_6) \wedge (p_{13} \vee p_{14}) \wedge (p_{15} \vee p_{16}) \wedge (p_{17} \vee p_{18})$

$\alpha_{3_{3|4}} = (\neg p_1 \vee p_2) \wedge (p_7 \vee \neg p_8) \wedge (p_9 \vee \neg p_{10}) \wedge (p_{13} \vee \neg p_{14}) \wedge (p_{15} \vee \neg p_{16}) \wedge (p_{17} \vee p_{18})$

$\alpha_{4_{3|4}} = (\neg p_3 \vee p_4) \wedge (\neg p_7 \vee p_8) \wedge (p_{11} \vee \neg p_{12}) \wedge (\neg p_{13} \vee p_{14}) \wedge (p_{17} \vee \neg p_{18}) \wedge (p_{17} \vee \neg p_{18})$

$\alpha_{5_{3|4}} = (\neg p_5 \vee p_6) \wedge (\neg p_9 \vee p_{10}) \wedge (\neg p_{11} \vee p_{12}) \wedge (\neg p_{15} \vee p_{16}) \wedge (\neg p_{17} \vee p_{18}) \wedge (\neg p_{17} \vee p_{18})$

Theorem 13: $\forall m > 1: \forall n > 1: \ell(\sum_{m|n}) = m \ \& \ d(\sum_{m|n}) = n$.

Proof of Theorem 13 (in two parts):

(1) It follows straightforwardly that $\forall m > 1: \forall n > 1: \mathcal{a}(\Sigma_{m|n}) = n$, since the instances of schema 4 which are used to construct $\Sigma_{m|n}$ are disjoint with respect to their atoms. Thus, $\forall a \subseteq \Sigma_{m|n}: |a| < n \Rightarrow a \nvdash_{PL} \bot$, and $\forall a \subseteq \Sigma_{m|n}: |a| = n \Rightarrow a \vdash_{PL} \bot$.

(2) Since $\forall a \subseteq \Sigma_{m|n}: |a| < n \Rightarrow a \nvdash_{PL} \bot$, and $\forall a \subseteq \Sigma_{m|n}: |a| = n \Rightarrow a \vdash_{PL} \bot$, the least size of a partition of $\Sigma_{m|n}$ which possess all consistent cells will be composed of m-1 (n-1)-membered cells and 1 1-membered cell.

Corollary 2: $\forall m > 1: \forall n > 1: \exists \Sigma: \ell(\Sigma) = m \ \& \ \mathcal{a}(\Sigma) = n$.

§2.2.2 Generalisations:

Having established the distinctness of the two measures it may be noted that generalisations of the two can be generated by varying the property for which we will measure the dilution or level of. Two generalisations, which enable us distinguish a denumerably infinite number of distinct measures of coherence, come to mind. We can speak of the level of a set for some dilution of incoherence, and we may also speak of the dilution of a set for levels of incoherence other than two.

Definition 8: For a set of sentences, Σ, the level of Σ for dilution n, $\ell_{[n]}(\Sigma)$, is the size of the least partition of Σ into subsets none of which possess a dilution less than n. (If there is no such partition, the level of a set for dilution n is an arbitrarily large value, ∞.)

Formally, $\ell_{\lfloor n \rceil}(\Sigma) = \min\{ m \mid \exists A \in \prod_m(\Sigma): \forall a \in A: \alpha(\Sigma) > n \}$ if $\perp \notin \Sigma$, and
$= \infty$ else.

Corollary 3: $\forall\Sigma: \ell_{\lfloor n \rceil}(\Sigma) = \ell(\Sigma)$.

An example:

Let $\Delta = \Sigma_{3|3} \cup \Sigma_{4|4}$ (where $\Sigma_{3|3}$ and $\Sigma_{4|4}$ are instances of schema 5,
and are assumed to be disjoint with respect to their atoms.)

In that case: $\ell(\Delta) = 4$, $\alpha(\Delta) = 3$, $\ell_{\lfloor 2 \rceil}(\Delta) = 1$, $\ell_{\lfloor 3 \rceil}(\Delta) = 3$, and $\ell_{\lfloor 4 \rceil}(\Delta) = 4$.

The following generality is evident:

Theorem 14: $\forall\Sigma: \alpha(\Sigma) > n \Rightarrow \forall s \geq n: \ell_{\lfloor s \rceil}(\Sigma) = 1$.

.

Definition 9: For a set of sentences, Σ, the dilution of Σ for level m, $\alpha_{\lfloor m \rceil}(\Sigma)$, is the size of the smallest subset of Σ of level m. (If the level of the set is less than m, the dilution of the set for level m is some arbitrarily large value, ∞.)[14]

More formally, $\alpha_{\lfloor m \rceil}(\Sigma) = \min\{ n \mid a \subseteq \Sigma \ \& \ |a| = n \ \& \ \ell(\Sigma) = m \}$, if $\ell(\Sigma) \geq m$, and
$= \infty$ else.[15]

Corollary 4: $\forall\Sigma: \alpha_{\lfloor 2 \rceil}(\Sigma) = \alpha(\Sigma)$.

An example: Let $\Delta = \{ p \wedge q \wedge r, \neg p \wedge q \wedge r, p \wedge \neg q \wedge r, \neg r \}$.

---

[14] Ray Jennings and Martin Allen introduced me to this generalisation.

[15] It is not surprising that the definition of corruption can be also generalised as follows:
$C_{\lfloor m \rceil}(\Sigma) = |\{ c \subseteq \Sigma \mid \ell(c) \geq m \}| \ / \ |\wp(\Sigma)|$.
Theorem 5 can also be generalised according to the generalisation of dilution for varied levels. Thus we have: **Theorem 5b:** $\forall\Sigma: \alpha_{\lfloor m \rceil}(\Sigma) \geq n \Rightarrow C_{\lfloor m \rceil}(\Sigma) \leq |\{ c \subseteq \Sigma \mid |c| \geq n \}| \ / \ |\wp(\Sigma)|$.

Then $\alpha_{\zeta[4]}(\Delta) = 4$, $\alpha_{\zeta[3]}(\Delta) = 3$, and $\alpha_{\zeta[2]}(\Delta) = 2$.

Given our new definition, the existence of a quite general relationship between our original definition is evident. This relationship can be codified by the following theorem.

Theorem 15: $\forall\Sigma: \forall n > 1: \alpha(\Sigma) \geq n \Rightarrow \alpha_{\zeta[m]}(\Sigma) \geq (m-1)(n-1)+1$.

Proof of Theorem 15: The proof is straightforward. If $\alpha(\Sigma) \geq n$, then the smallest possible subset of Σ of level m would have to be composed of m-1 (n-1)-membered cells, and 1 1-membered cell.

Having just considered a pair of generalisations of the measures *dilution* and *level* of incoherence, it is of interest to note that it seems that there must be an indenumerably infinite number measures of coherence which could be defined by iterating our generalisations. For example, we can measure a set's level for dilution n of level m (that is $\ell_{\alpha_{\zeta[n]}[\zeta[m]]}(\Sigma)$), a set's level for dilution n of level m for dilution o (that is, $\ell_{\alpha_{\zeta[n]}[\zeta[m]]_{\zeta[o]}}(\Sigma)$) etc. In the next section I explain the potential use of some generalisations of dilution of incoherence.

§2.2.2.1 A Problem with Dilution for the Proposed Application, and its Solution:

Having recognised some of the virtues of dilution as a measure of the degree to which we abide by the coherence principle, I now wish to consider one of its limitations.

Given the intended application, dilution appears to be a rather clumsy measure of the relative complexity of a set of sentences, where relative complexity is taken to encumber cognitive accessibility and the detection, and derivation of inconsistency. The generic variety of dilution appears to conflate complexity with cardinality. This obviously reflects a gross oversimplification of the cognitive phenomena whose representation is desired, as the following sets demonstrate:

$$a' = \{ (((p \vee (q \vee r)) \wedge (q \vee (r \vee \neg p))) \wedge (q \vee (r \supset p)) \wedge (q \vee (p \supset \neg r)),$$
$$(\neg q \vee (p \vee r)) \wedge (\neg(\neg(\neg(q \supset r) \supset \neg p) \vee \neg(\neg q \vee \neg r \vee p)) \wedge ((r \supset \neg q) \vee \neg p)) \}$$

$$b = \{ p, p \supset q, q \supset r, \neg r \}$$

The two sets illustrate that the size of a set is not an adequate measure of its complexity, for while the dilution of a' ($\mathcal{d}(a') = 2$) is less than the dilution of b, there are some grounds for saying that the inconsistency of b is easier to detect. The problem suggested by the present example does not undermine the claim that, as a consequence of the methodology of interpretation, we ought to maximise the dilution of the sets of intentional states we attribute. What the example does show, however, is that on occasion intuition indicates that some less dilute sets are easier to draw contradictions from than more dilute ones. This fact disturbs the apparent correspondence between the measure by which we ought to maximise the coherence of the sets of intentional we attribute, and the measure of an individual's ability to detect inconsistency among his set of intentional states.

A possible solution to the problem would be to concern ourselves with agent relative *dilution-profiles.* Such profiles would amount to placing restrictions on the sets of intentional states we would attribute, according various generalisations of dilution. For example, we could define various generalisations of dilution where inconsistency is defined as provability of a sentence of the form $\alpha \wedge \neg\alpha$ from a sound but incomplete codification of classical logic. Thus, for example, if we define the system $L_{PL3} = < L, A, R >$, where L is the language of propositional logic, and where A is all instances of the schema $(\neg\alpha \supset \neg\beta) \supset (\beta \supset \alpha)$, and where R = { [MP], [&I] }, we may define a correlate measure of dilution.

Definition 10: *The dilution of incoherence of* Σ for L_{PL3}, written $\mathcal{d}_{L_{PL3}}(\Sigma)$, is equal to the size of the smallest L_{PL3}-inconsistent subset of Σ. (If the set is L_{PL3}-consistent, the dilution for L_{PL3} of a set is some arbitrarily large value, ∞.)

With such a measure in hand we may stipulate the permissibility of attributions save where the dilution for L_{PL3} is less than 5. Such a restriction would permit us attribute a′, but not b.

Another possible (and it seems to me more practical) solution to the present problem would be to introduce operators for the purpose of marking intentional states (by kind or by the domains to which their content is relevant) for the purposes of placing a floor on the dilution of a set relative to varied marked types of states. According to this solution to the problem, we would attach operators to the elements of a′ to indicate their convoluted syntax. The strategy of marking states according to their type may have other uses. For example, it might be advantageous to formally

distinguish between beliefs and desires. We could thereby allow a person's set of desires to possess a lesser degree of dilution than his set of beliefs. This would be satisfying, for the attribution of self conscious inconsistent desires, ambivalence, seems more tolerable than the attribution of self conscious inconsistent beliefs, though both types of incoherence tend to undermine our ability to make principled predictions of an agent's behaviour.

§2.3 Coherence Maximisation as a Dictate of Interpretive Strategy:

Despite certain desirable features it seems that the measure, *dilution*, would not in itself, if used as a constraint on interpretation, enforce on the sets of intentional states we attribute, a characteristic which it seems any set of intentional states should possess. Any coherence measure employed as a constraint on interpretation should reflect not only the acuity and vigilance of reasoners, but also the fact that having true beliefs safeguards the existence of their possessors. Similarly, for social beings such as ourselves, having false beliefs invites correction and hence the extinction of the false beliefs. Contrary to the demand that such considerations should be reflected by our interpretive practice, a constraint on the dilution of the sets of intentional states we attribute does not rule out the attribution of sets of intentional states where all subsets of the set of size equal to a stipulated lower limit on dilution are inconsistent. One way of illustrating why this is problematic is to consider the close relationship between consistency and truth. The point is made using a measure of consistency which is designed to reflect the close relationship between truth and consistency.

Definition 11: The minimum false part of a set, Σ, MinF(Σ), is equal to the minimum number of sentences which must be removed from Σ to make it consistent.

Some Examples:

$$MinF(\{ p, \neg p\}) = 1.$$

$$MinF(\{ p, \neg p, r, \neg q \}) = 1.$$

$$MinF(\{ p, \neg p, q, \neg q \}) = 2.$$

Theorem 16a: $\forall n > 1: \forall \Sigma: \mathcal{d}(\Sigma) \geq n \Rightarrow MinF(\Sigma) \leq |\Sigma| - (n-1)$[16].

Theorem 16b: $\forall n > 1: \exists \Sigma: \mathcal{d}(\Sigma) = n \ \& \ MinF(\Sigma) = |\Sigma| - (n-1)$.

Proof of Theorems 16a and 16b: Theorem 16a is obviously true. Moreover, schema 5 illustrates how we would generate sets which would confirm 16b for any n.

The point made by theorem 16b is that if the size of a set exceeds its dilution by a substantial degree, then it may be that the set is terribly incoherent. For example, if the size of a set is twice its dilution, then it may be that most of the sentences in the set must be false, in virtue of formal properties alone!

The proposal for placing a floor on the dilution of the sets of intentional states we attribute to agents does not appear, in itself, to amount to a sufficient guarantee for the coherence of those sets of intentional states. It should be mentioned that

---

[16] Similarly, $\forall n > 1: \forall \Sigma: \mathcal{d}(\Sigma) \leq m \Rightarrow MinF(\Sigma) \leq |\Sigma| - \left\lceil |\Sigma| / m \right\rceil$.

despite my recognition of this problem, the considerations which now compel us to find (more) coherence in the sets of intentional states we attribute are different than the methodological considerations which counsel us to embrace the coherence principle. The consideration which compels us to abide by the coherence principle is an *a priori* assumption that in doing so we increase the likelihood that the sets of intentional states we attribute will support predictions of behaviour. On the other hand, the current concern for increasing the coherence of the sets of intentional states we attribute appears to be *strategic*. On the basis of assumptions about an agent's belief-generating processes and about the agent's history, we reason that the agent's set of intentional states ought to be mostly coherent and mostly true.

In the next two sections I will explore a seductive remedy to the problem reflected in theorem 16b. While this remedy does not turn out to be wholly satisfactory, it is relevant to the problem, and besides, the line of thought is of intrinsic interest.

§2.3.1 Consistency and Cardinality:

Given the fact that having true beliefs safeguards the existence of their possessors, we ought to suppose that any strategy which increases the likelihood of the truth of an agent's beliefs will tend to be adopted as a result of informed or natural selection. One strategy which indirectly promotes truth, by promoting coherence, is that of restricting the size in addition to the dilution of one's set of intentional states. Before suggesting several strategies for deriving a principled means for fixing a ceiling on the size of the sets of intentional states we attribute, it should be mentioned that many individuals, including Davidson, have argued that it does not make sense to

count beliefs.(Davidson 1983, p.308 & 1975, p.156-7) Davidson's claim is based on two considerations. Both have to do with the way he thinks propositions get their identity. According to Davidson, identifying a belief requires locating it in a space of logical relations which define its content and causal efficacy. For example, a belief cannot be the belief that it is raining, if it does not motivate one to open one's umbrella on an occasion when one wishes to keep dry.(Davidson 1985, p.351-2) Similarly, the possession of particular beliefs tends to implicate semantic and conceptual competence. Thus, for example, one cannot believe that a gun is loaded, if one does not believe that a gun is a weapon, etc.(Davidson 1975, p.156-7) The apparent consequence of these observations is that the possession of any one belief requires the possession of indefinitely many others.

Despite Davidson's warnings, it makes perfect sense to count the number of beliefs we attribute to an agent. As I stated above, my interest here is in discovering a measure suitable as a constraint on acceptable sets of basic explanatory intentional states. Moreover, given the definition of a set of basic explanatory intentional states, as a set of intentional states which are sufficient for explaining the behaviour of a subject up to the present , it is not inconsistent with Davidson's view to suppose the legitimacy of counting such intentional states.

§2.3.2 Preserving Level by Preserving Dilution:

Respecting statements made by Davidson to the effect that we should rationalise inconsistency by the attribution of cognitive partitions, I will now show that knowing the size of a set along with its dilution (for some level), permits us to infer an upper limit of the set's level. This fact demonstrates the benefit of constraining the size in addition to

the dilution of a set of intentional states, or of a body of information, generally. In effect, the following theorems have relevance to the task of reasoning about one's own cognitive shortcomings, since the conceptual tools employed in reasoning about the ignorance of others are equally applicable in the case of reasoning about one's own possible ignorance.

Theorem 17: $\forall\Sigma\colon \forall n > 1\colon \mathcal{d}(\Sigma) \geq n \Rightarrow \mathcal{l}(\Sigma) \leq \left\lceil {}^{|\Sigma|}/_{n-1} \right\rceil$.

Proof of Theorem 17: We want to know (given that the dilution of a set is equal to or greater than n) the maximum number of (mutually inconsistent) cells which could be used to construct a $|\Sigma|$-membered set. The proper strategy would to make the cells as small as possible. However, we must obey the prior assumption that all possible (n-1)-tuples of sentences will be consistent. Thus, we suppose that the maximum partition of Σ is composed of $|\Sigma|/(n-1)$ (n-1)-membered cells, plus one cell, if $|\Sigma|$ is not divisible by n-1.

Theorem 18:

$$\forall\Sigma\colon \forall n > 1\colon \forall m > 2\colon \mathcal{d}_{\{m\}}(\Sigma) \geq n \Rightarrow \mathcal{l}(\Sigma) \leq \left\lfloor \frac{|\Sigma| - \left(\mu \times \left\lfloor \frac{n}{m} \right\rfloor\right)}{\left\lceil \frac{n}{m} \right\rceil} \right\rfloor + \mu.$$

$$\mu = \left(\left\lceil \frac{n}{m} \right\rceil - \left(\frac{n}{m}\right)\right) \times m.$$

Proof of Theorem 18: Here, we want to know the maximum partition of a $|\Sigma|$-membered set, such that no m-tuple of cells has a union consisting of less than n members.

The general rule for maximising the size of such a partition is to assume that the union of the first m-cells has n members. Moreover, we should assume that those cells

distribute the first n elements into cells of the size $\lceil n/m \rceil$ and $\lfloor n/m \rfloor$. (There will be μ elements of size $\lfloor n/m \rfloor$.) The remaining $|\Sigma|$ - n elements will be divided among cells of size $\lceil n/m \rceil$, save the remaining elements which may placed any cell of the partition.

An example illustrates the point of why the first n sentences will be divided among cells of size $\lceil n/m \rceil$ and $\lfloor n/m \rfloor$:

Suppose that n=7 and m=4. In that case there will be 3 possible ways of dividing 7 sentences among 4 cells:

(a) $\{\{\alpha, \alpha, \alpha, \alpha\}, \{\alpha\}, \{\alpha\}, \{\alpha\}\}$

(b) $\{\{\alpha, \alpha, \alpha\}, \{\alpha, \alpha\}, \{\alpha\}, \{\alpha\}\}$

(c) $\{\{\alpha, \alpha\}, \{\alpha, \alpha\}, \{\alpha, \alpha\}, \{\alpha\}\}$

Looking at these possible partitions, we can note that if $|\Sigma|$ is 7, then it would make no difference which way we partitioned the set, the maximum partition of Σ would be 4. However, if we assume that Σ possesses more than 7 elements, then we notice that (c), the partition that distributes the 7 sentences most evenly, will always form a part of a partition of Σ consisting of the greatest possible number of cells. This follows from the fact that in adding new cells to our first m cells, each of the added cells must be as large as the largest cell of the first m cells, since any smaller cell could form a part of an m-tuple of cells whose union possessed fewer than n members.

Here are some examples of the pronouncements of the theorem:

IF $\alpha_{[3]}(\Sigma) = 5$ & $|\Sigma| = 17$, **THEN** $\ell(\Sigma) \leq 8$.

IF $\alpha_{[3]}(\Sigma) = 5$ & $|\Sigma| = 30$, **THEN** $\ell(\Sigma) \leq 15$.

IF $\alpha_{[3]}(\Sigma) = 14$ & $|\Sigma| = 30$, **THEN** $\ell(\Sigma) \leq 6$.

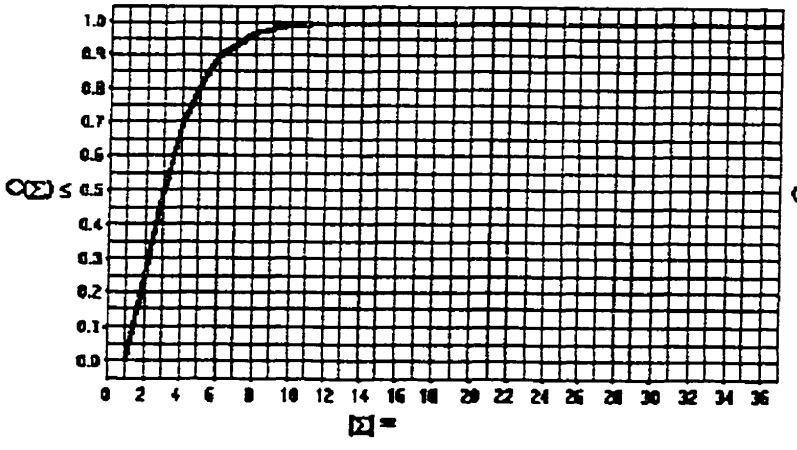
IF $\alpha_{[5]}(\Sigma) = 14$ & $|\Sigma| = 30$, **THEN** $\ell(\Sigma) \leq 10$.

One interesting consequence which follows from the existence of theorems 19 and 20 is that any inference which preserves a measure of dilution and cardinality will also preserve a degree of level of incoherence. For example, if **R** is a relation which may hold between a set, a sentence and a natural number, such that $\forall\Sigma, \alpha: \forall n > 1: < \Sigma, \alpha, n > \in \mathbf{R} \Leftrightarrow \mathcal{d}(\Sigma \cup \{\alpha\}) > n \ \& \ |\Sigma \cup \{\alpha\}| \leq 2n$, then it follows that $\forall\Sigma, \alpha: \forall n > 1: < \Sigma, \alpha, n > \in \mathbf{R} \Rightarrow \mathcal{d}(\Sigma \cup \{\alpha\}) < 3$.

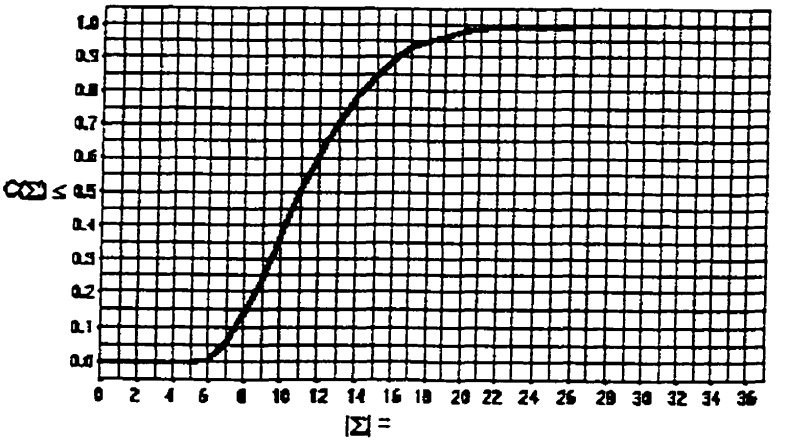
§2.3.3 Corruption Again:

The fact that a restriction on the size of a set in addition to a restriction on its dilution gives us a greater assurance of the set's coherence is demonstrated by the following graphs (which represent the entailments codified by theorem 5). The following graphs of an upper limit of $C(\Sigma)$ in comparison to $|\Sigma|$, where $\mathcal{d}(\Sigma)$ is fixed, demonstrate the manner in which a guarantee of greater dilution induces a greater resistance to increases in an upper limit of $C(\Sigma)$ as we add sentences to a set.

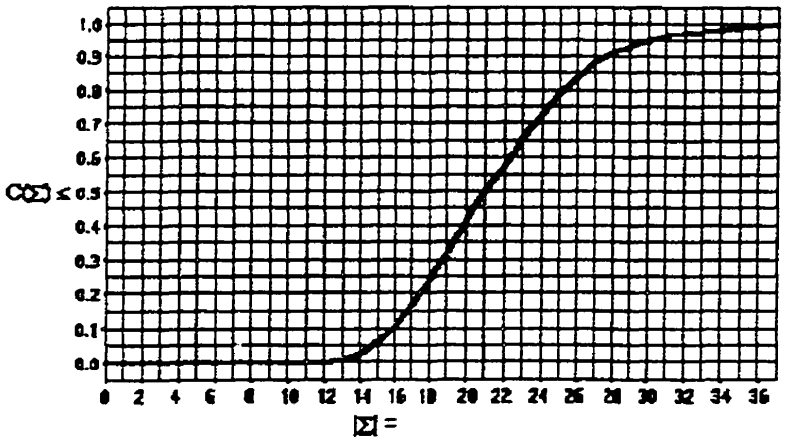
$\mathcal{d}(\Sigma) = 1$



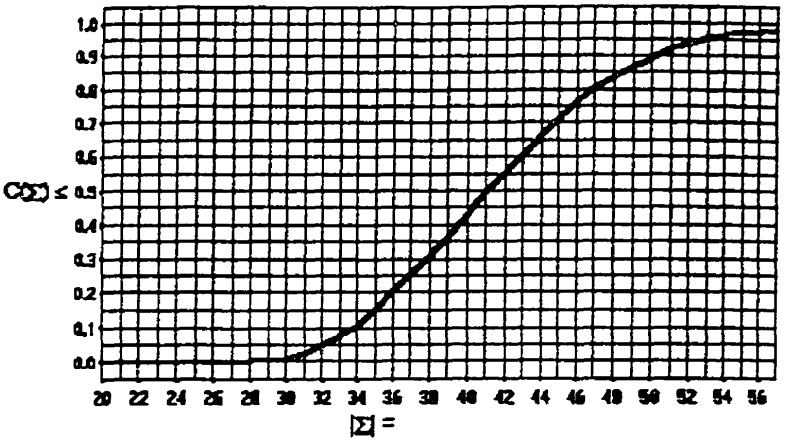
$\mathcal{d}(\Sigma) = 5$



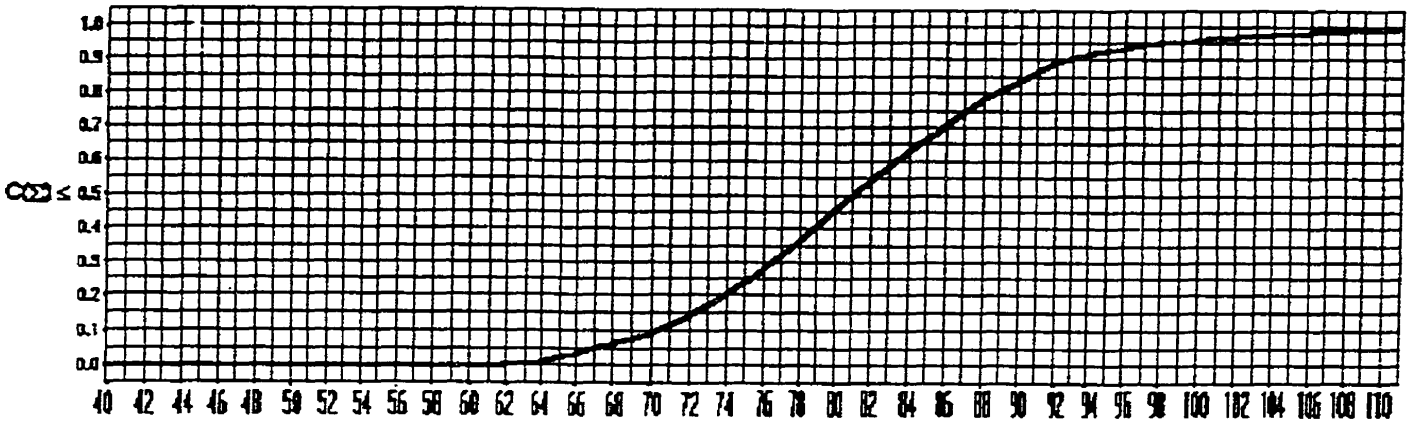
$\alpha(\Sigma) = 10$



$\alpha(\Sigma) = 20$



$\alpha(\Sigma) = 40$



We can take the information represented by the preceding graphs to be relevant to our reasoning about the degree to which we should restrict the quantity of intentional states which we attribute to an agent. Moreover, it may be useful to regard the preceding graphs as representing a relation between two cognitive magnitudes (one which is represented by $\alpha(\Sigma)$, and the other by the $|\Sigma|$). Indeed, if we shift our

interpretation slightly we can think of the elements of $\wp(\Sigma)$ as representing the potentially expressible propositions of some language and in turn we can think of the sentences which compose these sets as the language's set of propositional atoms. Finally, then, we can think of the value of $\mathcal{A}(\Sigma)$ as representing the degree of our intellectual ability to grasp the various propositions potentially expressible in our language. Given this fiction we can see that the preceding graphs give us a nice representation of the trade-off one encounters when one contemplates adding an additional sentence/atom to our language. For each sentence/atom we add to Σ we increase the size of our conceptual vocabulary, this is, $|\wp(\Sigma)| - |\{ c \subseteq \Sigma \mid |c| \geq n\}|$, while decreasing the likelihood that the set of our concepts is relatively uncorrupted. Interestingly, differences in the proportion of what is lost and gained varies greatly and systematically with the size of one's set of 'atoms' and with the value of $\mathcal{A}(\Sigma)$. Several patterns are evident.

For example, $\forall\Sigma, \alpha: \forall n > 1: <\Sigma, \alpha, n> \in \mathbf{R} \Rightarrow C(\Sigma) < 1/2$.[17]

§2.3.4 A Summary of Section 2.3:

I claimed earlier that part of the objective of chapter two was to lay out a sort of annotated menu of several coherence measures which could be applied as part of a method of interpretation. In spite of the many options I argued that the measure, *dilution of incoherence*, would be particularly useful. Despite the claim for the aptness of this measure, the observations made at the beginning of section 2.3 were supposed to show up the defect of attempting to grade the coherence of a set of

---

[17] In virtue of features of their mathematical basis, the respective upper limits of $C(\Sigma)$ (which are derived via theorem 5) can be displayed in a form which possesses the characteristic property of Pascal's triangle. For more on this, see appendix C.

intentional states using only the measure, *dilution.* What I went on to show was that restricting the size of a set along with its dilution tends to restrict the incoherence of the set.

It is likely that restricting the size of the sets of intentional states we attribute would be insufficient to reflect the impact which various means of selection are likely to have on the coherence of an agent's set of intentional states. This follows from the fact that a restriction on the size of the sets of intentional states we attribute would not entail that the coherence of such sets will reflect a statistical impact which would most likely be the result of selective forces. Moreover, a constraint on size in addition to dilution does not rule out the attribution of sets of intentional states where all subsets of the set of size equal to our stipulated floor on dilution are inconsistent. All that is ensured by restricting size in addition to dilution is that the set of such subsets is smaller than it might otherwise have been. And thus minimising the size of one's set of intentional states seems more likely to be a sort of strategy an agent would adopt given its beneficial effect.

Much of the burden of ensuring the coherence of attributed sets of intentional states not carried by the imposition of a constraint on the dilution of such sets will be carried by the probable use of heuristics which leave a positive mark on the sets of intentional states we attribute. For example, one of the most significant strategies which we no doubt use in interpretation is that of attributing clusters of intentional states to agents in cases where an agent falls under one or more sortals. We probably also attribute degrees of acuity and vigilance in the maintenance of a set of intentional states according to an agent's membership in a kind. The principal kinds of kinds we discern as an instrument of interpretation are such kinds as species,

occupation, and religious affiliation, etc. By attributing clusters according to these kinds we embrace a strategy which has consequences for the coherence of the sets of intentional states we attribute in virtue of the large degree of coherence possessed by many (though not all) of these clusters.

Despite recognising the effect of using the sort of heuristics just mentioned, we can also implement the strategy of directly restricting the incoherence of the sets of intentional states we attribute by measures other than dilution. If we pursue this strategy along with the strategy of restricting the size of the sets we attribute, we can measure the divergence of our restrictions with what is formally dictated by the mere restriction of size and dilution. Thus, if we restrict the level of incoherence of the sets we attribute in addition to their size and dilution we can measure the difference between the constraint we place on the set and the upper limit on the level of the set codified by theorem 20. Similarly, if we choose to restrict the dilution of the sets we attribute for some level other than (and in addition to) level two, we may measure the impact of such restrictions by measuring the difference between our restriction and what is formally dictated by any prior restriction on the set for dilution of level two (using theorem 15).[18]

I would now like to reemphasise the point that there are two distinct kinds of reasons for limiting the incoherence of the sets of intentional states we attribute. One kind of reason which suggests that we should limit the incoherence of the sets of intentional states we attribute follows from assumptions about how sets of

---

[18] The strategy of restricting the dilution of the sets of intentional states we attribute for levels other than two is appealing since it accommodates the fact that it is often the case that our interpretations of individuals are schematic. In most cases, it is not worthwhile to attribute a full-blown set of basic behaviour-explaining intentional states that could be counted.

intentional states tend to be tested by the world. Thus it is proposed that we recognise that 'the constraint of the world' tends to ensure that sets of intentional states are more likely to be coherent than incoherent. In turn, the adoption of presuppositions about the coherence of belief sets *as a result of the world* can be thought of as *strategic* in the way they have an impact upon the interpretive process. Assuming that the intentional states of an individual are largely coherent (and true) is a strategy which reflects our assumption that natural selection and the process of living and learning selects for coherence. Similarly, the strategy of restricting the size of the sets of intentional states we attribute reflects our assumption that natural selection and the process of living and learning selects for the adoption of coherence promoting strategies. These sorts of strategic considerations are different from the primary consideration which motivated my claim that we ought to maximise the coherence of the sets of intentional states we attribute. This consideration is not merely strategic, but rather methodological since at least part of the reason for adopting the principle is independent of any empirical assumptions about the histories of the individuals we interpret. Rather the principle is justified by the fact that in adhering to it we ensure that the intentional states we attribute will serve as a sound basis for *making predictions* of behaviour.

Chapter 3:

Logics of Belief Attribution.

§3.1 Historical Roots — Paraconsistency and the Preservationist Strategy:

Although it has many virtues, an obvious defect of the classical consequence relation is that it permits unprincipled inference from inconsistent sets of premisses. While this feature would seem to amount to a reason for rejecting the classical consequence relation, classical semanticists have construed the unattractive feature as merely a curious by-product of an essential feature of any acceptable consequence relation. This putative essential feature is that a consequence relation should never lead us from true premisses to a false conclusion. Thus, the apparent defect of the classical consequence relation is countenanced on the grounds that a classically unsatisfiable premiss set cannot be true, and thus trivial consequences could naturally be drawn on the basis of such an unsatisfiable state of affairs. Far from excusing the unwanted characteristic of the classical consequence relation, it seems that this reasoning merely rationalises its curious behaviour on correspondingly curious grounds. While the unwanted feature may be a by-product of the putatively essential property of any acceptable consequence relation, the question remains whether the unwanted feature is inescapable. If not, then there is no positive reason why it should not be purged. In fact there are positive reasons why it should: a logic that does permit principled inferences from inconsistent data ought to be welcomed (or at least not rejected out of hand), since the body of data we have at our disposal is so often inconsistent. Evidently our ordinary inferential practices are already non-classical.

Some logicians (R. and V. Routley in (1972), N. Belnap in (1976), G. Priest in (1979), and N. Rescher and R. Brandom in (1980), for example) have responded to the apparent defect of the classical consequence relation by developing semantics which employ valuations that permit the satisfaction of classically inconsistent premiss sets. Having done this it becomes possible to define a notion correlative to classical validity. Typically, in such systems, the inference from a premiss set Σ to a conclusion α is valid if and only if the satisfaction conditions for Σ are a superset of the set of satisfaction conditions for α. While this approach accomplishes the desired result, it does so at the cost of leaving the meaning of our logical vocabulary, particularly negation, virtually unrecognisable. A less drastic approach, that appears to uphold the truth-conditional meaning of our logical vocabulary (and is thus suitable for the present application), has been devised by Jennings and Schotch. The basis of their solution consists in the explicit recognition of what the putative essential feature of classical logic amounts to. This is merely that an acceptable consequence relation should preserve truth. Acting as good Samaritans, Jennings and Schotch have pointed out that the unwanted characteristic of the classical consequence relation is separable from the putative necessary feature of an acceptable consequence relation. They have demonstrated that the problem with the classical consequence relation can be remedied by merely augmenting the classical consequence relation so as to produce consequence relations that preserve properties (namely various measures of coherence) in addition to truth. This approach, which has been called the preservationist strategy,

has general applicability to the development of preservationist logics, in addition to paraconsistent logics.

§3.2 Formal Preliminaries:

A natural extension of the project of defining measures which can be used to measure the coherence of sets of basic explanatory intentional states, is to define formal systems which can guide principled inferences from an agent's possession of a set of basic explanatory intentional states to his possession of other intentional states. These other intentional states would be thought of as implicit in an agent's set of basic explanatory intentional states. Whether or not such formal systems exist, it is evident that it would be straightforward to define consequence relations which capture the permissible inferences.

Definition 12: A *consequence relation* is a set, *C*, of ordered pairs, $\langle \Sigma, \alpha \rangle$, each of which is composed of a premiss set and a sentence. The second element of a pair, α, is said to be a *consequence* of the first, Σ, according to *C*, if and only if $\langle \Sigma, \alpha \rangle \in C$. (I will also use the semantic notation $\Sigma \models_C \alpha$ to indicate that $\langle \Sigma, \alpha \rangle \in C$.)

If we know which property we would like to preserve, there are consequence relations which would capture those inferences which we ought to regard as permissible. We want consequence relations that preserve a selected property.

Definition 13: A *bivalent property*, ψ, *(of sets of sentences)* is a set of subsets of Φ (the set of well-formed formulae of a selected formal language).[1]

The following convention is also adopted: we say $\psi(\Sigma)$ if and only if $\Sigma \in \psi$,

---

[1] Throughout this essay I will use ψ as a variable ranging over bivalent properties of sets of sentences.

and also $\psi(\Sigma) = 1$, if $\psi(\Sigma)$, else $\psi(\Sigma) = 0$.

Definition 14: A consequence relation, *C*, *preserves ut nunc* a bivalent property, ψ, (written $PRES^{un}(C, \psi)$) if and only if for every pair in *C*, if the premiss set possesses ψ, then the union of the premiss set with its consequence possesses ψ.[2]

$$\text{Formally, } PRES^{un}(C, \psi) \Leftrightarrow (\forall\Sigma,\alpha: \langle\Sigma, \alpha\rangle \in C \Rightarrow (\psi(\Sigma) \Rightarrow \psi(\Sigma\cup\{\alpha\}))).$$

Given this definition, it is easy to see how, for any bivalent property, we could define respective consequence relations which would preserve it. We could use the following schema:

Schema 6: $\forall\Sigma,\alpha: \langle\Sigma, \alpha\rangle \in C_\psi \Leftrightarrow (\psi(\Sigma) \Rightarrow \psi(\Sigma\cup\{\alpha\}))$.

While this result may be interesting, it is not of much use, for reasons which will be elaborated later. Moreover, what we would prefer to have for the present application are axiomatic or deductive formal systems which preserve the selected measures of coherence. That is, we would like to have formal systems which are sound and complete with respect to a consequence relation which preserves the selected measures of coherence. We want to possess a formal system, rather than merely a consequence relation (which preserves a desired measure of coherence), since we want to be able to apply a codifiable set of inferential procedures. The consequence relation with which such a set of procedures is sound and complete, will insure that the procedures we apply meet predetermined standards of adequacy (namely, for the proposed application, that the inferential procedures preserve a selected degree of coherence).

---

[2] This definition is preliminary.

A formal system, S, is an ordered triple, < L, A, R >, consisting of a language, L, a (possibly empty) set of axioms, A , and a set of rules, R. Moreover, given a formal system, we can define a notion of proof.

Definition 15: There is a proof (in S) of α from Σ, written $\Sigma \vdash_s \alpha$, if and only if there is a finite sequence of sentences, $\beta_1, \ldots, \beta_n$, (where $\beta_n = \alpha$), and where each β_i is either, (1) an axiom of S, (2) a sentence of Σ, or (3) the output of a rule of S.

We can now express a preliminary form of our desiderata.

We want (1) a consequence relation, C^*, such that at least $PRES^{un}(C^*, \psi)$, where ψ is a degree of coherence, and (2) a formal system, *S*, such that

$\forall \Sigma, \alpha$: $\Sigma \models_{C^*} \alpha \Leftrightarrow \Sigma \vdash_s \alpha$.

§3.3 Level Preservation:

Part of the idea behind the level-preserving consequence relation defined by Jennings and Schotch, called *forcing*, is simply to restrict the aggregation of data which comes from potentially mutually inconsistent sources. Forcing accomplishes this effect by permitting classical inference tempered by a restriction on level increasing aggregation.

Definition 16: A set, Σ, *n-forces* a sentence, α, which we write $\Sigma \lceil\models_n \alpha$, if and only if every partition of Σ into n cells includes at least one cell that classically implies α.

The consequence relation defined by $\ell(\Sigma)$-forcing can be called *Level preserving consequence relation* (or LPCR).

Definition 17: $\langle \Sigma, \alpha \rangle \in$ LPCR, which we write $\Sigma \left[\models_{\ell(\Sigma)} \alpha$, if and only if every partition of Σ into $\ell(\Sigma)$ cells includes at least one cell that classically implies α.

The formal deductive system consisting of the following rules was proved to be sound and complete for n-forcing (See Apostoli and Brown (1995)):

$$[\text{Ref}]: \quad \frac{\alpha \in \Sigma}{\Sigma \left[\vdash_n \alpha}$$

$$[\text{Pres}^* \mid\!\sim_{PL}]: \quad \frac{\Sigma \left[\vdash_n \alpha \;\&\; \alpha \vdash_{PL} \beta}{\Sigma \left[\vdash_n \beta} \qquad [\text{Trans}]: \quad \frac{\Sigma, \alpha \left[\vdash_n \beta \;\&\; \Sigma \left[\vdash_n \alpha}{\Sigma \left[\vdash_n \beta}$$

$$[2/n+1 \text{ - INTRO}]: \quad \frac{\Sigma \left[\vdash_n \alpha_1 \;\&\; \ldots \;\&\; \Sigma \left[\vdash_n \alpha_{n+1}}{\Sigma \left[\vdash_n \underbrace{((\alpha_1 \wedge \alpha_2) \vee \ldots \vee (\alpha_n \wedge \alpha_{n+1}))}_{\text{all 2-member conjunctions from } \{\alpha_1, \ldots, \alpha_{n+1}\}}}$$

§3.3.1 Fixed and Floating Forcing:

Both the consequence relation, *forcing*, and its associated deductive system, *the logic of forcing*, can be used to preserve level in two ways. More generally, for any

property of sets which admits of degrees, we can speak of consequence relations which preserve a fixed or floating measure of the property.

Definition 18: A *measure of a property*, φ, *(of sets of sentences) which admits of degrees* is the output of a function φ: $\wp(\Phi) \rightarrow \{ N \cup \{\infty\} \}$, where N is some set of numbers, and ∞ an arbitrary object.[3]

There are several particularly useful bivalent properties which can be associated with each measure of a property, φ, which admits of degrees. Thus, the following conventions are also adopted:

$$\varphi^{[n]}(\Sigma) \text{ if and only if } \Sigma \in \varphi^{[n]} \text{ if and only if } \varphi(\Sigma) = n.$$
$$\varphi^{\lceil n \rceil}(\Sigma) \text{ if and only if } \Sigma \in \varphi^{\lceil n \rceil} \text{ if and only if } \varphi(\Sigma) \leq n.$$
$$\varphi^{\lfloor n \rfloor}(\Sigma) \text{ if and only if } \Sigma \in \varphi^{\lfloor n \rfloor} \text{ if and only if } \varphi(\Sigma) \geq n.$$

Definition 19: A consequence relation, *C*, *preserves ut nunc a fixed measure, n*, of a property, φ (which admits of degrees), if and only if $PRES^{un}(C, \varphi^{[n]})$, or $PRES^{un}(C, \varphi^{\lceil n \rceil})$, or $PRES^{un}(C, \varphi^{\lfloor n \rfloor})$.

Thus, for example,

$$PRES^{un}(C, \varphi^{[n]}) \Leftrightarrow (\forall \Sigma, \alpha: \langle \Sigma, \alpha \rangle \in C \Rightarrow (\varphi^{[n]}(\Sigma) \Rightarrow \varphi^{[n]}(\Sigma \cup \{\alpha\}))).$$

Definition 20: A consequence relation, *C*, *preserves ut nunc a floating measure* of a property, φ (which admits of degrees) if and only if $\forall \Sigma$: $PRES^{un}(C, \varphi^{[\varphi(\Sigma)]})$.

---

[3] Throughout this essay I will use φ as a variable ranging over non-bivalent properties of sets of sentences.

Moreover,

$\forall\Delta$:PRES$^{un}(C, \varphi^{[\varphi(\Delta)]}) \Leftrightarrow (\forall\Sigma,\alpha: <\Sigma, \alpha> \in C \Rightarrow (\varphi^{[\varphi(\Sigma)]}(\Sigma) \Rightarrow \varphi^{[\varphi(\Sigma)]}(\Sigma\cup\{\alpha\}))).$[4]

In these terms, we can express some of the characteristics of the consequence relations defined by Jennings and Schotch.

Theorem 19: $\forall$n: PRES$^{un}([\models_n , \mathcal{C}^{n}])$. **Corollary 5:** $\forall\Sigma$: PRES$^{un}([\models_{\mathcal{A}(\Sigma)}, \mathcal{C}^{[\mathcal{A}(\Sigma)]}).$[5]

§3.3.2 Monotonicity:

There are a family of notions deserving the label monotonicity. Indeed, the term monotonicity can be applied variously to forms of reasoning and to consequence relations.

Definition 21: A consequence relation, $\models_c$, is *monotonic* if and only if

$$\forall\Sigma, \Delta, \alpha: \Sigma \models_c \alpha \Rightarrow \Sigma \cup \Delta \models_c \alpha.$$

Theorem 20: $\forall$n:$[\models_n$ is monotonic. **Theorem 21:** $\forall\Sigma$: $[\models_{\mathcal{A}(\Sigma)}$ is not monotonic.

We can generally think of properties of sets as being monotonic or non-monotonic, and we can characterise monotonicity as a sort of inferential stability under informational increase.

---

[4] This definition maintains the form of the general definition of preservation. But, of course, $\forall\Sigma, \varphi: \varphi^{[\varphi(\Sigma)]}(\Sigma)$.

[5] These statements do not adequately express Schotch and Jennings' results. Additionally, the consequence relation they defined has the property *preserving level within closure.* This property will be described in the following section.

Definition 22: A bivalent property, ψ, is *naturally monotonic* if and only if

$$\forall\Sigma: \Sigma \in \psi \Rightarrow \forall \Sigma^{*}: \Sigma \subset \Sigma^{*} \Rightarrow \Sigma^{*} \in \psi.$$

For example, inconsistency is naturally monotonic. On the other hand, consistency is naturally non-monotonic (that is, not naturally monotonic).

We can also distinguish between the stability of a set's possession of a property upon the addition of new information, and stability of a set's possession of a property under the addition of consequences derived from the set. This distinction gives rise to a natural generalisation of the notion of monotonicity as it is applied to consequence relations. Thinking of the monotonicity of a consequence relation as its stability under the addition of information, we can make a distinction between the addition of genuinely new information, and the addition of consequences that can be extracted from sets by the consequence relation. Expressing this fact compels me to introduce a general definition of closure, as it applies to consequence relations:

Definition 23: The *n^{th} degree closure* of a set, Σ, under a consequence relation, C, is the set, $CL_{\vdash_c}^{[n]}(\Sigma) = \{ \Sigma^{n} \mid (\exists c \subseteq \Sigma : c \models_{C} \alpha) \Rightarrow (\alpha \in \Sigma^{1} \ \& \ \alpha \in \Sigma^{n}) \ \&$
$(\exists c \subseteq \Sigma \cup \Sigma^{1} : c \models_{C} \alpha) \Rightarrow (\alpha \in \Sigma^{2} \ \& \ \alpha \in \Sigma^{n}) \ \&...$
$...\& (\exists c \subseteq \Sigma \cup \Sigma^{1} \cup ... \cup \Sigma^{n-1} : c \models_{C} \alpha) \Rightarrow \alpha \in \Sigma^{n} \}.$

(For all Σ, let $CL_{\vdash_c}^{[0]}(\Sigma) =_{Dfn} \varnothing$.)

The expression $CL_{\vdash_c}(\Sigma)$ will be used as an abbreviation of $CL_{\vdash_c}^{[\infty]}(\Sigma)$.

Given the notion of n-degree closure, we can define a correlative notion of monotonicity.

Definition 24: C is *Monotonic within n^{th} degree closure* (i.e., C is $MON^{[n]}$) if and only if $\forall\Sigma, \Delta, \alpha: (\Sigma \models_c \alpha \ \& \ \Delta \subseteq CL_{\models_c}{}^{[n]}(\Sigma)) \Rightarrow \Sigma \cup \Delta \models_c \alpha$.

A welcome feature of the consequence relation that Jennings and Schotch designed is that the set of consequences from a set licensed by forcing is stable under use. Indeed, both of the ways of forcing I mentioned above have the property of being $MON^{[\infty]}$. (I will refer to any consequence relation that is $MON^{[\infty]}$ as *monotonic within closure*.)

Theorem 22: $\forall\models_c$: $\models_c$ is monotonic, then $\models_c$ is monotonic within closure.

That forcing is monotonic within closure is a virtue, since a consequence relation which preserves level of incoherence need not be, consider $C_{\ell[\mathcal{A}\Sigma]}$:

Definition 25: $\forall\Sigma, \alpha: \Sigma \models_{C_{\ell[\mathcal{A}\Sigma]}} \alpha \Leftrightarrow (\ell^{[\ell(\Sigma)]}(\Sigma) \Rightarrow \ell^{[\ell(\Sigma)]}(\Sigma \cup \{\alpha\}))$.[6]

Theorem 23: $C_{\ell[\mathcal{A}\Sigma]}$ is not monotonic within closure. That is:

$$\exists\Sigma, \Delta, \beta: \Sigma \models_{C_{\ell[\mathcal{A}\Sigma]}} \beta \ \& \ \Delta \subseteq CL_{\models_{C_{\ell}[\mathcal{A}\Sigma]}}(\Sigma) \ \& \ \text{Not}(\Sigma \cup \Delta \models_{C_{\ell[\mathcal{A}\Sigma]}} \beta).$$

Proof of Theorem 23 (by example):

Where $\Sigma = \{ p \wedge q, \neg p \wedge q, p \wedge \neg q, r \wedge s, \neg r \wedge \neg s \}$

$\Sigma \models_{C_{\ell[\mathcal{A}\Sigma]}} r \wedge \neg s \ \& \ \Sigma \models_{C_{\ell[\mathcal{A}\Sigma]}} \neg r \wedge s$. But not $\Sigma \cup \{ r \wedge \neg s \} \models_{C_{\ell[\mathcal{A}\Sigma]}} \neg r \wedge s$.

---

[6] This Consequence relation is an instance of schema 6.

While being non-monotonic is a unobjectionable property for a consequence relation slated for a preservationist application, not being monotonic within closure may be problematic. We can, of course, prune consequence relations, according to the following general schema:

$\forall\Sigma, \Delta, \alpha$: $n = \max\{ n \mid C \text{ is } MON^{[n]} \} \Rightarrow \langle\Delta, \alpha\rangle \in C^{\Sigma\text{-PRUNED}} \Leftrightarrow \langle\Delta, \alpha\rangle \in CL_{\vdash_c}^{[n]}(\Sigma)$.

Theorem 24: $\forall\Sigma$, C: $C^{\Sigma\text{-PRUNED}}$ is $MON^{[\infty]}$.

Despite this fact, it is clear that the rules or theorems of many consequence relations which are not monotonic within closure cannot be usefully codified. More importantly, the usefulness of any consequence relation which is not monotonic within closure will probably be limited as a basis for drawing consequences which preserve some selected property, since even if we build a clause into our consequence relation stipulating that it preserve some selected property, there is no guarantee that it will do so to a satisfactory degree. The addition of any pair of individually property preserving consequences to a set may result in our failing to preserve the property. This was shown in the case of $C_{\ell[\alpha\Sigma]}$. We need a more general definition of preservation sufficient to express our wish to avoid consequence relations such as $C_{\ell[\alpha\Sigma]}$.

Definition 14′: A consequence relation, *C*, *preserves* a bivalent property, ψ, to the n^{th} degree of closure (written $PRES^{[n]}(C, \psi)$) if and only if

$$\forall\Sigma: \psi(\Sigma) \Rightarrow \psi(CL_{\vdash_c}^{[n]}(\Sigma)).$$

The expression $PRES(C, \psi)$ will be used to abbreviate $PRES^{[\infty]}(C, \psi)$

I will say that a consequence relation, *C*, *preserves* a bivalent property, ψ, within closure if and only if $PRES(C, \psi)$.

Schotch and Jennings' results can now be stated properly:

Theorem 26: $\forall n$: PRES($[\models_n, \mathcal{C}^{[n]}]$). **Corollary 6:** $\forall \Sigma$: PRES($[\models_{\alpha(\Sigma)}, \mathcal{C}^{[\alpha(\Sigma)]}]$).

It is also evident that we may define consequence relations that preserve any selected property within closure. We can use instances of the following schema:

$\langle \Sigma, \alpha \rangle \in C_{\psi} \Leftrightarrow$
$\langle \Sigma, \alpha \rangle \in \cap\{ \models_C \mid \text{PRES}(\models_C, \psi) \ \& \ \forall \models_D : \text{PRES}(\models_D, \psi) \Rightarrow \models_C \subset \models_D \}$.

Definition 14′ also gives us the resource to state other results.

For example:

Definition 26: $\langle \Sigma, \alpha \rangle \in C_{n\text{-AG}} \Leftrightarrow \alpha = \beta_1 \wedge \ldots \wedge \beta_n \ \& \ \{ \beta_1, \ldots, \beta_n \} \subseteq \Sigma$.

Theorem 27: $\forall \Sigma$: $\forall s$: $\text{PRES}^{[1]}(C_{(\lceil \alpha(\Sigma)/(s-1) \rceil - 1)\text{-AG}}, \mathcal{C}^{[s]})$.

Theorem 28: $\forall \Sigma$: $\alpha(\Sigma) \geq n \Rightarrow \text{PRES}^{[\lceil \log_2 n \rceil + 1]}(C_{2\text{-AG}}, \mathcal{C}^{[2]})$.

We can make even finer discriminations by adopting the following definition:

Definition 14″: A consequence relation, *C*, *preserves* a bivalent property, ψ, to the n^{th} degree for m (written $\text{PRES}^{[n,m]}(C, \psi)$) if and only if

$\forall \Sigma$: $\forall \alpha_1, \ldots, \alpha_m \in CL_{\models_C}^{[n]}(\Sigma)$: $\psi(\Sigma) \Rightarrow \psi(CL_{\models_C}^{[n-1]}(\Sigma) \cup \{ \alpha_1, \ldots, \alpha_m \})$.

Given this definition the following equivalence can be observed:

Theorem 25: $\forall C, \psi$: $\text{PRES}^{[1,1]}(C, \psi) \Leftrightarrow \text{PRES}^{un}(C, \psi)$.

§3.3.3 Other Level-preserving Strategies:

In choosing a consequence relation that preserves level, LPCR is certainly not the only available choice. If we imagine a continuum of the most conservative to the most liberal consequence relations which preserve level of incoherence it is evident that LPCR is somewhere in the middle. We can think of a consequence relation that permits no aggregation as the most conservative, and on the other hand $C_{\zeta[\Phi\Sigma]}$ is obviously the most liberal. We can also imagine an ordering of consequence relations according to the proper superset relation. I will say that a consequence relation, C_1, is stronger than another, C_2, if and only if $\models_{c2} \subset \models_{c1}$.

Obviously, not all pairs of consequence relations are comparable in this way, since it may be for two consequence relations that neither is a proper superset of the other. However, it is useful to rate comparable consequence relations by strength, since for any application it would generally be preferable to have the strongest consequence relation possible (that preserves some selected property). For instance, LPCR is obviously not the strongest consequence relation which preserves level of incoherence. For example, $C_{\zeta[\Phi\Sigma]}$ is a stronger. But, of course, $C_{\zeta[\Phi\Sigma]}$ does not preserve level of incoherence within closure and is therefore not really a competitor of LPCR. There are, however, other consequence relations which are stronger than LPCR, and preserve level within closure. I will now describe one, with the help of a few of definitions.

Definition 27: $\alpha \in \Sigma^{BASE}$ if and only if $\alpha \in \Sigma$, and there are no non-equivalent elements of Σ from which α can be classically proved.

In other words: $\Sigma^{BASE} = \{ \alpha \in \Sigma \mid \forall \beta \in \Sigma: (\{\beta\} \vdash_{PL} \alpha) \Rightarrow (\{\alpha\} \vdash_{PL} \beta) \}$.

Definition 28: $\alpha \in \Sigma^{INRT}$ if and only if $\alpha \in \Sigma^{BASE}$, and the union of α and any consistent subset of Σ^{BASE} is also consistent.
In other words: $\Sigma^{INRT} = \{ \alpha \in \Sigma^{BASE} \mid \forall b \subseteq \Sigma^{BASE}: b \nvdash_{PL} \bot \Rightarrow b \cup \{\alpha\} \nvdash_{PL} \bot \}$.

Definition 29: $AUG(\Sigma) = \{ \alpha = \beta \wedge \gamma_1 \wedge ... \wedge \gamma_n \mid \beta \in \Sigma \ \& \ \Sigma^{INRT} = \{ \gamma_1, ..., \gamma_n \} \}$.

Given these definitions, a consequence relation stronger than LPCR can be defined.

Definition 30: A set, Σ, *n-forces*$^+$ a sentence, α, which we write $\Sigma \ [\models^+_n \alpha$, if and only if α follows classically from the union of Σ^{INRT} and at least one cell of every n-partition of Σ.

Definition 31: $\langle \Sigma, \alpha \rangle \in LPCR^+$, which we write $\Sigma \ [\models^+_{\ell(\Sigma)} \alpha$, if an only if α follows classically from the union of Σ^{INRT} and at least one cell of every partition of Σ into $\ell(\Sigma)$ cells.

Theorem 29: $\forall n: PRES(\ [\models^+_n, \ell^{[n]})$.

Proof of Theorem 29:
The result follows from the fact that:
(1) $\forall \Sigma: \ell(\Sigma) = \ell(\Sigma^{BASE})$;
(2) $\forall \Sigma: \ell(\Sigma^{BASE}) = \ell(AUG(\Sigma^{BASE}))$;
(3) $\forall \Sigma: \ell(AUG(\Sigma^{BASE})) = \ell(AUG(\Sigma))$; and
(4) $\forall \Sigma, n: CL_{[\models^+_n}(\Sigma) = CL_{[\models_n}(AUG(\Sigma))$.

Corollary 7: $\forall\Sigma$: PRES($[\models^{+}_{\mathcal{A}\Sigma}, \mathcal{L}^{\mathcal{L}(\Sigma)}]$).

The possibility of codifying LPCR^{+} or another consequence relation stronger than LPCR which preserves level invites further research, since in choosing a level-preserving strategy there are grounds for choosing a formal system which is sound and complete with respect to a stronger consequence relation which preserves level within closure over a weaker one.[7] Given our aim of discovering logic which preserves dilution, the moral of discovering a stronger relation applies. In fact, the moral is particularly applicable, given the proposed application, for as we saw in some of the examples is section 2.1.4, it may be that none of an agent's potentially active sets of intentional states possess enough content relevant to a situation to enable the agent to decide how to act. It may be that the problem in such cases could be remedied, if we were able to make strong inferences about what implicit (but discernible) intentional states we may suppose the agent to possess.

§3.4 The General Applicability of Forcing:

Since its inception the importance of level of incoherence as a measure of coherence has been recognised. To a large degree this is because the measure and the consequence relation that preserves it, *forcing*, are generalisations of classical consistency, and the classical consequence relation, respectively. Another appealing characteristic of the measure was the identification of the logic which corresponds to forcing with a class of modal logics weaker than K, so-called weakly-

---

[7] There is also an intrinsic interest in understanding the structure of all of the MON$^{[\sim]}$ consequence relations which preserve level of incoherence.

aggregative modal logics.[8] Despite the apparent virtues of the measure itself, it is arguable that it is not the measure itself that is of great significance, rather it is the consequence relation, *forcing*, and its logic that are most significant. The measure, level of incoherence, derives its preeminence from the directness of its connexion with the logic of forcing.

A significant feature of the logic of forcing lies in its potential use as a system of deduction that preserves properties other than level of incoherence. In fact, in general, we can ask of a formal system *S* (or consequence relation *C*): what properties does *S* (or *C*) preserve? It is evident the logic of forcing will preserve many generalisations of the measure, *level*. The measure, *level*, can be generalised in ways in addition to the ones mentioned in chapter two.

Definition 32: For a set of sentences, Σ, the level of Σ for ψ, $\ell_\psi(\Sigma)$, is the size of the least partition of Σ into subsets none of which possess ψ. (If there is no such partition, the level of a set for ψ is an arbitrarily large value, ∞.)

Given this definition, the following general preservational property of forcing can be stated:

Theorem 30: $\forall\psi$: ($\forall\Sigma$: $\Sigma \not\in \psi \Rightarrow CL_{\models_n}(\Sigma) \not\in \psi$) $\Rightarrow$ $\forall n$: PRES($[\models_n, \ell_\psi^{\lceil n \rceil})$.

Despite this general result, it is evident that there are bivalent properties, ψ, such that n-forcing does not preserve $\ell_\psi^{\lceil n \rceil}$. For example, n-forcing does not preserve

---

[8] The frame theory for these logics is a natural generalisation of the kripkean frame theory. (See Schotch and Jennings (1980).)

$\zeta_{\mathcal{L}[\Sigma]}^{[n]}$. Just consider the set $\Delta = \{ p \wedge r, (p \supset q) \wedge r, \neg q \wedge r, \neg r \}$. $\zeta_{\mathcal{L}[\Sigma]}(\Delta) = 2$. However, Δ 2-forces $p \wedge (p \supset q) \wedge r$. Despite the limitations of the correspondence between forcing and generalisations of the measure, *level of incoherence*, it is likely that forcing at some value will preserve $\zeta_{\mathcal{L}[\Sigma]}^{[n]}$. Moreover, for any naturally non-monotonic property it makes sense to ask: what value of forcing will preserve it?

The reason why forcing can be put to use preserving properties other than level of incoherence can be made evident. n-Forcing preserves level of incoherence n for the reason that if a set, Σ, is of level n, then there will be a partition, π, of Σ into n cells, each of which is classically consistent. Moreover, given such a partition of Σ, the union of the classical closure of each of π's cells taken individually will also be of level n. In turn, the union of the classical closure of each of π's cells will be a superset of the set of consequences which follow from every n-partition of Σ.

To discover a logic of forcing that preserves some other property, it is necessary only to discover a method for determining the size, n, of the smallest partition of a set such that we could take the union of the classical closure of each of the cells of the partition individually without generating a set that varies according to the chosen property. n-Forcing from the set will then preserve this property. Some definitions will make this point clearer.

Definition 33: The *restricted classical closure* of a family A of sets, written Rc(A), is the set $\{ \alpha \mid \exists a \in A: a \vdash_{PL} \alpha \}$.

And, of course, a level-preserving partition of a set, Σ, is a partition, $\pi \in \Pi(\Sigma)$, such that $\ell^{[\ell(\Sigma)]}(\mathrm{Rc}(\pi))$.[9]

Definition 34: The set of level-preserving partitions of a set, Σ, is the set:

$$\Pi^{[\ell(\Sigma)]}(\Sigma) = \{ \pi \in \Pi(\Sigma) \mid \ell^{[\ell(\Sigma)]}(\mathrm{Rc}(\pi)) \}.$$

This definition can be generalised as follows:

Definition 35: For any property, φ, which admits of degrees the set of φ-preserving partitions of a set, Σ, is the set: $\Pi^{\varphi^{[\varphi(\Sigma)]}}(\Sigma) = \{ \pi \in \Pi(\Sigma) \mid \varphi^{[\varphi(\Sigma)]}(\mathrm{Rc}(\pi)) \}$.

(I will adopt the notation of writing $\Pi^{\varphi^*}(\Sigma)$ to express $\Pi^{\varphi^{[\varphi(\Sigma)]}}(\Sigma)$.)

More general still we have:

Definition 36: For any bivalent property, ψ, the set of ψ-preserving partitions of a set, Σ, is the set: $\Pi^{\psi}(\Sigma) = \{ \pi \in \Pi(\Sigma) \mid \psi(\mathrm{Rc}(\pi)) \}$.

Definition 37: The *pseudo-level* for level-preservation, $\ell^{[\ell(\Sigma)]}(\Sigma)$, is the size of the smallest member of $\Pi^{[\ell(\Sigma)]}(\Sigma)$, if $\Pi^{[\ell(\Sigma)]}(\Sigma) \neq \varnothing$, else $\ell^{[\ell(\Sigma)]}(\Sigma) = \infty$.

That is, $\ell^{[\ell(\Sigma)]}(\Sigma) = \min \{ |\pi| \mid \pi \in \Pi^{[\ell(\Sigma)]}(\Sigma) \}$, if $\Pi^{[\ell(\Sigma)]}(\Sigma) \neq \varnothing$.
$= \infty$ else.

This definition is generalised as follows:

---

[9] Note that $\forall \Sigma, \pi$: $\ell^{[\ell(\Sigma)]}(\mathrm{Rc}(\pi))$ *if and only if* $\ell(\mathrm{Rc}(\pi)) = \ell(\Sigma)$.

Definition 38: The *pseudo-level* for φ-preservation, $\ell^{\varphi^{[\varphi(\Sigma)]}}(\Sigma)$, is the size of the smallest member of $\prod^{\varphi^{[\varphi(\Sigma)]}}(\Sigma)$, if $\prod^{\varphi^{[\varphi(\Sigma)]}}(\Sigma) \neq \varnothing$, else $\ell^{\varphi^{[\varphi(\Sigma)]}}(\Sigma) = \infty$.

(I will adopt the notation of writing $\ell^{\varphi^*}(\Sigma)$ to express $\ell^{\varphi^{[\varphi(\Sigma)]}}(\Sigma)$).

More general still we have:

Definition 39: For any bivalent property, ψ, the *pseudo-level* for ψ-preservation, $\ell^{\Psi}(\Sigma)$, is the size of the smallest member of $\prod^{\Psi}(\Sigma)$, if $\prod^{\Psi}(\Sigma) \neq \varnothing$, else $\ell^{\Psi}(\Sigma) = \infty$.

One of the distinctions of the measure, *level of incoherence*, lies in the fact that $\forall\Sigma: \ell(\Sigma) = \ell^{\alpha^*}(\Sigma)$. On the other hand, it is not true that $\forall\Sigma: \alpha(\Sigma) = \ell^{\alpha^*}(\Sigma)$, where, of course, $\ell^{\alpha^*}(\Sigma) = \min \{ |\pi| \mid \pi \in \prod^{\alpha^*}(\Sigma) \}$, and where

$\prod^{\alpha^*}(\Sigma) = \{ \pi \in \prod(\Sigma) \mid \alpha^{[\alpha(\Sigma)]}(\mathrm{Rc}(\pi)) \}$.

For example:
Where $\Gamma = \{ (p \wedge (\neg r_1 \vee \neg s_1) \wedge (r_2 \vee \neg s_2) \wedge (\neg r_1 \vee \neg r_2)), ((\neg p \vee q) \wedge s_1 \wedge s_2), (\neg q \wedge s_1 \wedge s_2), r_1, r_2 \}$, $\alpha(\Gamma) = 3$. However, the aggregation of any pair of elements of Γ will decrease dilution. Thus, $\ell^{\alpha^*}(\Gamma) = |\Gamma| = 5$.

The following generalisation does hold:

Theorem 31: $\forall\Sigma: \alpha(\Sigma) \leq \ell^{\alpha^*}(\Sigma)$.

The term *pseudo-level* is adopted for the reason that one might fail to distinguish the concept of level from the concept of pseudo-level, if one were preoccupied only with measuring and preserving the level of sets, since the measures are equivalent in this case. The marking out of the property designated by the expression is also welcomed since the property constitutes a more general measure which can be married to forcing and its logic. Indeed, where the pseudo-level for some property of a set is n, n-forcing from the set will preserve the property.

Theorem 32: $\forall\psi$: $\forall\Sigma$: PRES($\lceil\models_{\ell^{\psi}(\Sigma)}$, ψ).

Proof of Theorem 32:

$\forall\Sigma$: $\ell^{\psi}(\Sigma) = n \Rightarrow \exists\pi \in \prod_n(\Sigma)$: $\psi(Rc(\pi))$.

Therefore, $\forall\Sigma$: $CL_{\lceil\models_{\ell^{\psi}(\Sigma)}}(\Sigma) = CL_{\lceil\models_n}(\Sigma) \subseteq Rc(\pi)$. [By dfn. of $\lceil\models_n$, and $Rc(\pi)$.]

Despite the apparent hegemony of pseudo-level, it is clear that incoherence level has this virtue: level of incoherence equals pseudo-level for $\ell^{\lceil\ell(\Sigma)\rceil}$. Where such an identity does not hold for a property, φ, which admits of degrees, we forfeit a luxury: we cannot n-force from a set and thereby know that we have preserved $\varphi^{\lceil n\rceil}$ (or $\varphi^{\lfloor n\rfloor}$, depending on the property).

§3.5 Dilution Preservation:

The assumption behind the present project is that individuals possess preselected capacities to detect inconsistency among their own belief sets. This capacity is measured by the least size of an inconsistent set of beliefs that may escape the

person's notice. As interpreters we are urged to abide by a correlate restriction on the attribution of implicit intentional states, which embodies the supposition that no self-conscious inconsistencies are believed.

Definition 40: *Dilution preserving consequence relation* (DPCR)*:* $\langle \Sigma, \alpha \rangle \in$ DPCR, which we write $\Sigma \left[\models_{\ell^{\prec^*}(\Sigma)}\right. \alpha$, if and only if every partition of Σ into $\ell^{\prec^*}(\Sigma)$ cells includes at least one cell that classically implies α.

Corollary 8: $\forall \Sigma$: PRES($\left[\models_{\ell^{\prec^*}(\Sigma)}\right.$, $\mathcal{A}^{[\prec(\Sigma)]}$).

This result is rather uninformative. If we want to find out which logic of forcing preserves the dilution of a set, then we must find a property that is definitive of members of $\prod^{\prec^*}(\Sigma)$. Indeed, if we could find a property which is definitive of members of $\prod^{\prec^*}(\Sigma)$, this would permit us to discover the size of the smallest member of this set. In knowing the size of the least member of $\prod^{\prec^*}(\Sigma)$, we would thereby know the value at which we may force from a set and preserve its dilution.

The problem of discovering a property that defines members of $\prod^{\prec^*}(\Sigma)$ is slightly more complicated than discovering a property that defines members of $\prod^{\ell^*}(\Sigma)$ because we require that the cells of a dilution preserving partition possess some property in addition to classical consistency.

The significant feature of the property in question, is that, unlike consistency, it is context-sensitive. That is, the properties which a cell of a dilution preserving partition of a set may possess are constrained by the composition of the set itself and also by the proposed partition of the set of which the cell is an element. The characteristic context-sensitivity of the property suggests that we not attempt to look for a property that the individual cells of an acceptable partition must possess, but rather that we look for a property that a partition of such a set must possess. In fact, the property characteristic of members of $\prod^{\prec^*}(\Sigma)$ can be expressed.

The expression of this property compels us to enlist the following definition:

Definition 41: A is subsumed by B (written A $[\subseteq]$ B) if and only if

$\exists f: A \rightarrow B$ satisfying:

$\forall a, a^* \in A$: (i) $a \subseteq f(a)$, and

(ii) $a \neq a^* \Rightarrow f(a) \cap f(a^*) = \varnothing$.[10]

Theorem 33: $\forall \Sigma: \prod^{\prec^*}(\Sigma) = \{ \pi \in \prod(\Sigma) \mid \forall a \subseteq \Sigma: a \vdash_{PL} \perp \Rightarrow \exists \pi^* \in \prod_{\lhd\Sigma\rhd^+}(a): \pi^* [\subseteq] \pi \}$[11].

Proof of Theorem 33 (in two parts):

We prove:

[1] $\pi \notin \{ \pi \in \prod(\Sigma) \mid \forall a \subseteq \Sigma: a \vdash_{PL} \perp \Rightarrow \exists \pi^* \in \prod_{\lhd\Sigma\rhd^+}(a): \pi^* [\subseteq] \pi \} \Rightarrow \pi \notin \prod^{\prec^*}(\Sigma)$, &

[2] $\pi \in \{ \pi \in \prod(\Sigma) \mid \forall a \subseteq \Sigma: a \vdash_{PL} \perp \Rightarrow \exists \pi^* \in \prod_{\lhd\Sigma\rhd^+}(a): \pi^* [\subseteq] \pi \} \Rightarrow \pi \in \prod^{\prec^*}(\Sigma)$.

---

[10] In other words, A is subsumed by B *if and only if*

$\forall a \in A: \exists b \in B: a \subseteq b$, & $\forall a \& a^* \in A: \forall b \in B: a \subseteq b \& a \neq a^* \Rightarrow a^* \cap b = \varnothing$.

[11] Recall **Definition 2b:** $\prod_n(\Sigma)$ is the set of partitions of Σ whose cardinality is greater than or equal to n.

Proof of [1].

We assume that $\pi \in \prod(\Sigma)$, $\exists a \subseteq \Sigma$: a $\vdash_{PL} \perp$ and $\forall \pi^* \in \prod_{\mathcal{A}(\Sigma)^+}(a)$: $\sim(\pi^* [\subseteq] \pi)$.

(We prove that $\mathcal{A}(Rc(\pi)) < \mathcal{A}(\Sigma)$).

From our definition of subsumption we know that

$$\forall \pi^* \in \prod(a), |\pi^*| \geq \mathcal{A}(\Sigma): \quad Not(\forall c^* \in \pi^*: \exists c \in \pi: c^* \subseteq c)$$

Or

$$Not(\forall d^* \,\&\, c^* \in \pi^* \,\&\, \forall c \in \pi: c^* \subseteq c \,\&\, d^* \neq c^* \Rightarrow d^* \cap c = \varnothing).$$

That is, $\forall \pi^* \in \prod(a), |\pi^*| \geq \mathcal{A}(\Sigma): \quad \exists c^* \in \pi^*: \forall c \in \pi: c^* \not\subseteq c$

Or

$$\exists d^* \,\&\, c^* \in \pi^* \,\&\, \exists c \in \pi: c^* \subseteq c \,\&\, d^* \neq c^* \Rightarrow d^* \cap c \neq \varnothing.$$

We must prove that in either case $\mathcal{A}(Rc(\pi)) < \mathcal{A}(\Sigma)$.

Case 1:

Assume $\pi \in \prod(\Sigma)$, $\exists a \subseteq \Sigma$: a $\vdash_{PL} \perp$ and $\forall \pi^* \in \prod_{\mathcal{A}(\Sigma)^+}(a)$: $\exists c^* \in \pi^*$: $\forall c \in \pi$: $c^* \not\subseteq c$.

If for all $m \geq \mathcal{A}(\Sigma)$ there is no m-partition, π^*, of a such that every element of π^* is a subset of some cell of π, then either ($a \not\subseteq \Sigma$) OR ($|\pi| < \mathcal{A}(\Sigma)$).

But by our hypothesis $a \subseteq \Sigma$. Thus, $|\pi| < \mathcal{A}(\Sigma)$. It follows that $\mathcal{A}(Rc(\pi)) < \mathcal{A}(\Sigma)$.

Case 2: Assume $\pi \in \prod(\Sigma)$, $\exists a \subseteq \Sigma$: a $\vdash_{PL} \perp$ and $\forall \pi^* \in \prod_{\mathcal{A}(\Sigma)^+}(a)$:

$\exists d^* \,\&\, c^* \in \pi^* \,\&\, \exists c \in \pi: c^* \subseteq c \,\&\, d^* \neq c^* \Rightarrow d^* \cap c \neq \varnothing$.

It follows that $\exists a \subseteq \Sigma$: a $\vdash_{PL} \perp$ and $\forall \pi^* \in \prod_{\mathcal{A}(\Sigma)^+}(a)$:

$\exists d^* \,\&\, c^* \in \pi^* \,\&\, \exists c \in \pi: c^* \subseteq c \,\&\, d^* \neq c^* \Rightarrow d^* \cap c \neq \varnothing$.

If this is the case, then $|\pi| < |\pi^*|$.

But $|\pi^*| = \mathcal{A}(\Sigma)$.

It follows that $|\pi| < \mathcal{d}(\Sigma)$.

And thus $\mathcal{d}(Rc(\pi)) < \mathcal{d}(\Sigma)$.

Proof of [2]. Assume that $\mathcal{d}(Rc(\pi)) < \mathcal{d}(\Sigma) = n$.

Then $\exists A \subseteq \pi$, $A = \{ c_{i_1}, \ldots, c_{i_{n-1}} \}$: $\cup A \vdash_{PL} \bot$.

But it follows from our condition on π that $\forall A \subseteq \pi$: $A = \{ c_{i_1}, \ldots, c_{i_{n-1}} \} \Rightarrow \cup A \nvdash_{PL} \bot$.

(Indeed, for every inconsistent subset of Σ, π subsumes at least a $\mathcal{d}(\Sigma)$-partition of that set.)

Therefore, $\mathcal{d}(Rc(\pi)) = \mathcal{d}(\Sigma)$.

This concludes the proof.

§3.5.1 The Relationship between Level and Pseudo-Level for Dilution Preservation:

Given what we have seen so far (given schema 5, for example) it would be surprising if there did not turn out to be an entailment between a set's level and its pseudo-level for dilution preservation. Unfortunately, while the measure of what is entailed appears to be substantial in one direction, it is unclear how to prove the result. Surprisingly, in the other direction there is no entailment. First consider the apparent entailment between a set's level to its pseudo-level.

Conjecture: $\forall\Sigma{:}\forall m,n > 1$: $\ell(\Sigma) = m$ & $\mathcal{d}(\Sigma) = n \Rightarrow \ell^{\prec}(\Sigma) \geq (m-1)(n-1)+1$.

The absence of any entailment from a set's possession of a particular pseudo level for dilution preservation to the set's possession of a particular level of incoherence is illustrated by the following theorem.

Theorem 34: $\forall n > 2, \forall p \geq n: \exists \Sigma: \mathcal{d}(\Sigma) = n \ \& \ \ell^{\mathcal{d}^*}(\Sigma) = p \ \& \ \ell(\Sigma) = 2.$[12]

§3.5.2 A Generalisation:

In such cases where for some property φ, which admits of degrees, $\exists \Sigma: \varphi(\Sigma) \neq \ell^{\varphi^*}(\Sigma)$, it is clear that our ability to apply of the logic of forcing to the preservation of a measure of that property is somewhat hampered, since we can never n-force from a set at some value and thereby know that we have preserved $\varphi^{\lceil n \rceil}$ (or $\varphi^{\lfloor n \rfloor}$, depending on the property). We can, however, define pseudo-levels for a sort of fixed forcing by altering the minimum partition of the subsumed inconsistent subsets.

Definition 42: *m-Dilution preserving consequence relation* ($DPCR^{[m]}$): $\langle \Sigma, \alpha \rangle \in DPCR^{[m]}$, which we write $\Sigma \left[\models_{\ell^{\mathcal{d}^{[m]}}(\Sigma)} \alpha$, if and only if every partition of Σ into $\ell^{\mathcal{d}^{[m]}}(\Sigma)$ cells includes at least one cell that classically implies α.

Theorem 35:

$$\forall \Sigma: \forall m: \prod^{\mathcal{d}^{[m]}}(\Sigma) = \{ \pi \in \prod(\Sigma) \mid \forall a \subseteq \Sigma: a \vdash_{PL} \perp \Rightarrow \exists \pi^* \in \prod_{m*}(a): \pi^* [\subseteq] \pi \}.$$

Again, the proof of the equivalence is similar to the proof of theorem 28.

Corollary 9: $\forall \Sigma: \forall m: PRES(\left[\models_{\ell^{\mathcal{d}^{[m]}}(\Sigma)}, \mathcal{d}^{[m]} \right)$.

The following equivalence is also evident:

Theorem 36: $\forall \Sigma: \ell^{\mathcal{d}^{[2]}}(\Sigma) = \ell(\Sigma)$.

---

[12] The validity of the theorem follows from a schema for generating level 2 sets of any dilution greater than 1 and any pseudo-level for dilution preservation. The Schema is presented in Appendix D.

§3.5.4 Forcing for Greater Effect:

So far it has been shown that when $\ell^{\prec^*}(\Sigma) = n$ we may n-force from Σ and preserve $\mathscr{C}^{[\mathscr{C}\Sigma]}$. I have also claimed that in general when we are attempting to find a consequence relation which preserves some desired property we ought to choose a stronger consequence relation over a weaker one. Given this precept we might wonder whether or not it is possible to force at some value less than $\ell^{\prec^*}(\Sigma)$ and still preserve $\mathscr{C}^{[\mathscr{C}\Sigma]}$. In fact, we can. To reach the final result a series of lemmas must first be established.

Definition 43: When $d = \{ \alpha_1,...,\alpha_s \}$ (where $s \geq 1$) is a set of sentences, let $Ar(d) = \{ \alpha_1 \wedge ... \wedge \alpha_s \}$. Ar(d) will be called the arch-consequence of d.

Lemma 1:

$$\forall\Sigma: \forall n, p > 1: \ \ell^{\prec^*}(\Sigma) = p \ \& \ \mathscr{C}(\Sigma) = n \ \Rightarrow \exists\pi \in \prod_p(\Sigma): \forall b \subseteq \pi: |b| < n \ \Rightarrow \ \cup b \nvdash_{PL} \bot.$$

Proof of Lemma 1:

Assume for *reductio* that

$$\exists\Sigma: \ \ell^{\prec^*}(\Sigma) = p \ \& \ \mathscr{C}(\Sigma) = n \ \ \& \ \forall\pi \in \prod_p(\Sigma): \exists b \subseteq \pi: |b| < n \ \ \& \ \cup b \ \vdash_{PL} \bot.$$

By the definition of $\ell^{\prec^*}(\Sigma)$ we know that there is an element of the set $\prod^{\prec^*}(\Sigma) = \{ \pi \in \prod(\Sigma) \mid \mathscr{C}^{[\mathscr{C}(\Sigma)]}(Rc(\pi)) \}$ which has p elements.

Let π^* be a p membered element of $\prod^{\prec^*}(\Sigma)$.

But since, by our assumption, $\forall \pi \in \prod_p(\Sigma)$: $\exists b \subseteq \pi$: $|b| < n$ & $\cup b \vdash_{PL} \perp$,
it follows that $\exists b \subseteq \pi^*$: $|b| < n$ & $\cup b \vdash_{PL} \perp$.

But if $\exists b \subseteq \pi^*$: $|b| < n$ & $\cup b \vdash_{PL} \perp$, then $\mathcal{A}(Rc(b)) < n$.

Indeed, if $m < n$ & $b = \{c_1,...,c_m\}$ & $\cup b \vdash_{PL} \perp$, then $Ar(c_1) \cup ... \cup Ar(c_m) \vdash_{PL} \perp$.

But if $Ar(c_1) \cup ... \cup Ar(c_m) \vdash_{PL} \perp$, then $\mathcal{A}(Rc(b)) < n$.

And in that case $\pi^* \notin \{ \pi \in \prod(\Sigma) \mid \mathcal{A}^{[\mathcal{A}(\Sigma)]}(Rc(\pi)) \}$, contrary to hypothesis.

It follows from Lemma 1 that if $\ell^{\mathcal{A}^*}(\Sigma) = p$ and $\mathcal{A}(\Sigma) = n$, and if $\pi \in \prod_p(\Sigma)$ possesses the property mentioned in the consequent of lemma 1, then we can represent a useful subset of the set of models for the cells of π (and the arch-consequences of these cells). The models for these cells will be represented in the form of a matrix, $\tau(\Sigma)$.

When $\ell^{\mathcal{A}^*}(\Sigma) = p$ & $\mathcal{A}(\Sigma) = n$, $\tau(\Sigma)$ has the following composition:

(1) $\tau(\Sigma)$ has p columns (each is associated with a cell of π).

(2) $\tau(\Sigma)$ has $\binom{p}{n-1}$ rows (each is associated with an different (n-1)-membered subset of π). (Note that $|\{ c \mid c \subseteq \pi \& |c| = n\text{-}1 \}| = \binom{p}{n-1}$.)

(3) Each place of $\tau(\Sigma)$ is an intersection of a row and column. For each *place* in $\tau(\Sigma)$, there is a 1 in the place *only if* the place is in the intersection of a row and column, where the cell associated with the column is an element of the set associated with the row. 0's are in all other places of $\tau(\Sigma)$.

For example:

$$\forall\Sigma: \ell^{\prec^*}(\Sigma) = 4 \;\&\; \alpha(\Sigma) = 3 \Rightarrow \tau(\Sigma) =$$

		c_1	c_2	c_3	c_4	c_5
c_1,	c_2	1	1	0	0	0
c_1,	c_3	1	0	1	0	0
c_1,	c_4	1	0	0	1	0
c_1,	c_5	1	0	0	0	1
c_2,	c_3	0	1	1	0	0
c_2,	c_4	0	1	0	1	0
c_2,	c_5	0	1	0	0	1
c_3,	c_4	0	0	1	1	0
c_3,	c_5	0	0	1	0	1
c_4,	c_5	0	0	0	1	1

Lemma 2: Each row of $\tau(\Sigma)$ represents a model (we know to exist) in which every element of the respective cells of π are satisfied when there is a 1 in the column associated with the cell.

Proof of Lemma 2:

The theorem follows from the way $\tau(\Sigma)$ is composed and from the fact that if π is an element of the set $\prod^{\prec^*}(\Sigma)$ which has p elements, then $\forall b \subseteq \pi: |b| < n \Rightarrow b \not\vdash_{PL} \perp$, and thus $\forall b \subseteq \pi: |b| < n$, there is a model which satisfies b.

Given this Lemma, I will adopt the convention of saying that the cell associated with a column of $\tau(\Sigma)$ is satisfied in a model represented by a row of $\tau(\Sigma)$ if and only if the column has a 1 in the row.

$\tau(\Sigma)$ will eventually prove useful in establishing that n-forcing where n is less than $\ell^{*}(\Sigma)$ preserves dilution in some cases. As a step in that direction, the following few definitions are introduced.

Definition 44: $\tau^{m}(\Sigma)$ is simply $\tau(\Sigma)$ expanded by a set of columns whose composition of 1's is the set permutations of m 1's per column.

For example: $\forall\Sigma: \ell^{*}(\Sigma) = 4 \;\&\; \alpha(\Sigma) = 3 \Rightarrow$

$$\tau(\Sigma) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \;\&\; \tau^{4}(\Sigma) = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

In addition to representing a set of models for arch-consequences of cells of π, we can think of $\tau^{m}(\Sigma)$ as representing a set of models for elements of a set, Δ (where the elements of Δ are associated with the columns added to $\tau(\Sigma)$). Moreover, if for all (n-1)-tuples of sentences associated with the columns of $\tau^{m}(\Sigma)$, there exists a row of $\tau^{m}(\Sigma)$ which satisfies the (n-1)-tuple, we know that $\alpha(\Delta \cup \Sigma) = \alpha(\Sigma)$. (Note that this is the case in the preceding example.)

Definition 45: $g(\Sigma)$ = min g | for each (n-1)-tuple of sentences associated with the columns of $\tau^g(\Sigma)$ (including the arch-consequences of cells of π, where π is a dilution preserving partition of size $\ell^{\prec^*}(\Sigma)$), there exists a row of $\tau^g(\Sigma)$ which satisfies the n-1 tuple.

The following lemma is somewhat academic, for reasons which become clear in the proof of Theorem 33. However, the result may be relevant to future research. (The proof is in appendix E.)

Lemma 3: $\forall\Sigma$: $\forall$p,n >1: $\ell^{\prec^*}(\Sigma) = p \;\&\; \mathcal{A}(\Sigma) = n \Rightarrow g(\Sigma) = \binom{p}{n-1} - p + n - 1$.

Definition 46: A sentence β will be called $\tau(\Sigma)$-*stable* if and only if β is satisfied in at least $g(\Sigma)$ of the models represented by the rows of $\tau(\Sigma)$.

Definition 47: $s(\Sigma)$ is min n | for all n-tuples of arch-consequences of elements of π (where π is a dilution preserving partition of size $\ell^{\prec^*}(\Sigma)$), [2/n - INTRO] on the n-tuple outputs a $\tau(\Sigma)$-stable β.

Lemma 4: $\forall\Sigma$: PRES($[\models_{s(\Sigma)-1}, \mathcal{A}^{[\prec(\Sigma)]}$).

Proof of Lemma 4:

We need only observe why it is that [2/s(Σ) - INTRO] upon a set, Σ, will never permit a set of inferences whose dilution is less than $\mathcal{A}(\Sigma)$.

Since a truth functional compound β, formed by the disjunction of the pairwise conjunction of $s(\Sigma)$ elements of Σ, will either be satisfied in at least $g(\Sigma)$ of the models represented by $\tau(\Sigma)$ or will follow from a single cell of a set π^* (where $|\pi^*| = \ell^{\prec^*}(\Sigma)$ & $\pi^* \in \prod^{\prec^*}(\Sigma)$), we are guaranteed that $(s(\Sigma)-1)$-forcing will preserve $\alpha^{[\prec\Omega]}$.

An Observation: Any $s(\Sigma)$ columns of $\tau(\Sigma)$ has the property that at least $g(\Sigma)$ rows in the truncated matrix (which consists of $s(\Sigma)$ columns) have at least two 1's. (This holds in virtue of the truth function: $2/s(\Sigma)$.)

Finally, we are ready to establish a substantial result.

Theorem 37: $\forall\Sigma$: $\forall p,n>1$: $\ell^{\prec^*}(\Sigma) = p$ & $\alpha(\Sigma) = n \Rightarrow s(\Sigma) = p-n+3$.

Proof of Theorem 37:
We need only prove that it is sufficient to take p-n+3 columns of $\tau(\Sigma)$, if we wish to have at least $g(\Sigma)$ rows in this subset of the columns of $\tau(\Sigma)$ have at least two 1's.

This follows, since in taking p-n+3 columns of $\tau(\Sigma)$ we are assured that *every* row of our p-n+3 columned subset of $\tau(\Sigma)$ will have at least two 1's.

Indeed, each row of $\tau(\Sigma)$ has exactly p-n+1 0's.

Thus, by taking p-n+3 columns we are assured that each row of our p-n+3 columned subset of $\tau(\Sigma)$ will have two or more 1's, since assuming that all of the 0's in $\tau(\Sigma)$ are

in the chosen p-n+3 columned subset of $\tau(\Sigma)$, there still must be (p-n+3)-(p-n+1) = 2 1's per row in the p-n+3 columned subset of $\tau(\Sigma)$.

It follows that $\forall\Sigma$: $\forall p,n>1$: $\ell^{*}(\Sigma) = p$ & $\alpha(\Sigma) = n \Rightarrow s(\Sigma) = p-n+3$.

Corollary 10: $\forall\Sigma$: PRES($[\models_{\ell^{*}(\Sigma)-\alpha(\Sigma)+2}$, $\alpha^{[\alpha(\Sigma)]}$).

Corollary 10 represents a substantial improvement over corollary 8. Moreover, a further result can also be demonstrated:

Theorem 38: $\forall p, n>1: \exists\Sigma: \ell^{*}(\Sigma) = p$ & $\alpha(\Sigma) = n$ & $\alpha(CL_{[\models_{(p-n+1)}}(\Sigma)) < n$.

Proof 38: We need only consider elements of the set of sets such that if the dilution of a set, Σ, is n, then $\forall b \subseteq \Sigma$: $|b| = n \Rightarrow b \vdash_{PL} \perp$. (Schema 5 illustrates how to construct such sets.)
For all such sets $\ell^{*}(\Sigma) = |\Sigma|$. Moreover, for all such sets we know that there is no model which satisfies more than n-1 elements of the set.
In virtue of these properties, we know that for such sets $\tau(\Sigma)$ will be a well suited representation of a subset of the set of models for elements of Σ.
Indeed, no model will satisfy any superset of elements of Σ, which are not satisfied by a model represented by a row of $\tau(\Sigma)$.
Thus, the only non-equivalent models for elements of Σ which $\tau(\Sigma)$ does not represent are ones which satisfy less than n-1 elements of Σ.
Now suppose that we perform 2/p-n+2 introduction on the first p-n+2 elements of a set, Σ, which is an element of the set of sets just described, to form a sentence, β. (I am supposing that these p-n+2 elements of Σ are associated with the first p-n+2 columns of $\tau(\Sigma)$.)

We prove, in that case, that the remaining elements of $\tau(\Sigma)$ will have a cardinality of n-2, and will not be satisfied in any model for β.
Indeed, the remaining elements have a cardinality of p-(p-n+2) = n-2.
Yet any model for β is a model for as least two of the first p-n+2 elements of Σ (which are associated with the first p-n+2 columns of $\tau(\Sigma)$). [By the truth function: 2/p-n+2.]
Thus, at most (n-1)-2 = n-3 of the remaining elements are satisfied in any model which is a model for β.
Therefore, $\mathcal{d}(\Sigma \cup \{\beta\}) < n$.
This concludes the proof.

Despite this result, I should emphasise that [$2/(\ell^{\prec^{\cdot}}(\Sigma) - \mathcal{d}(\Sigma) + 3)$ - INTRO] on (distinct) elements of the set of arch-consequences of cells of a $\ell^{\prec^{\cdot}}(\Sigma)$ membered dilution preserving partition of a set invariably generates a set which is satisfied in each of the set of models represented by $\tau(\Sigma)$. One advantage of this is that the introduction rule preserves level of incoherence, as well as pseudo-level for dilution preservation. However, one would like to know whether there are stronger introduction principles that generate those consequences satisfied in exactly $g(\Sigma)$ rows of $\tau(\Sigma)$. Furthermore, one would like to know whether there is an algorithm for generating truth functions which do just this.

§3.6 Concluding Remarks:

As I mentioned in section 3.4.3, it may be that for the present application there are very good reasons for closing sets of intentional states under stronger rather than weaker dilution preserving logics. Given this concern a final moral

relevant to the methodology of interpretation becomes evident: if we intend to use a floor on dilution as a constraint on interpretation, we should also minimise the pseudo-level for dilution preservation of the sets of intentional states we attribute.

The conclusion of this chapter marks the end of my discussion of how one might go about formalising a method of interpretation which abides by the methodological assumptions outlined in chapter 1, and developed to some degree in chapter 2. Admittedly, the account of the proposed method is incomplete. Nevertheless, I believe that I have covered the important parts of the ground which is relevant to understanding the normative character of mental explanation, conceived broadly as a assumption concerning the rationality of the objects of interpretation. In fact, chapter 3 has gone beyond covering such ground, since I have set out some of the technical difficulties one would face if one were to attempt to implement the proposed method. I now return to matters more philosophical. Before doing so I would like to acknowledge that the lines of argument I adopt in chapter 4 seem to be against the spirit of the implicitly behaviourist approach to interpretation sketched in chapters 2 and 3. I will return toward the end of chapter 4 to explain the apparent shift in approach and explain my promiscuous estimation of the nature of interpretation. Now to the question of whether the mental is reducible to the physical.

Chapter 4:

The Prospects for Intentional Psychology.

In this chapter I challenge several views concerning the future and possible futures of intentional psychology. To this end, I will point out what I take to be defects in Davidson's arguments for the conclusion that, in principle, the mental is not reducible to the physical. And at the same time, I will articulate a view that draws on aspects of the Davidsonian picture, but accommodates the possibility of *some sort* of reduction of the mental to the physical. To conclude this essay I will touch upon the issue of eliminativism, and then provide a summary of some of the observations that have been made regarding the nature of intentional psychology. I begin the chapter by considering what are probably the two strongest arguments for the irreducibility of the mental.

§4.1 Two Arguments for the Irreducibility of the Mental:

If we accept either of the two following definitions of reduction, then the considerations that follow will give us good reason to believe that the mental is not reducible to the physical. *Definitional reduction* consists in the definition of the types of states (or predicates) of a coarser grained theory in the terms of the theory to which it is being reduced. *Nomological reduction* consists in the deducibility of the laws of a coarser grained theory from the laws of the theory to which it is being reduced through the application of a set of *bridge laws* which link a subset of the coarser grained theory's predicates to predicates of the theory to which it is being reduced. The two considerations are (a) that mental states are multiply realisable

and (b) that the contents of at least some mental states are determined by their external relations. If mental states are multiply realisable, then it is likely that any mental state can be realised by too many types of physical states for us to codify the physical application conditions for mental predicates. If the contents of mental states are determined by their external relations, then it is likely that for any class of states that we can define (according to the physical properties of its members), there will be pairs of members of the class which realise dissimilar mental state types.

Given my avowed intention to moderate Davidson's arguments for the irreducibility of the mental, and my claim that the two preceding considerations give us reason for concluding that the mental is not reducible to the physical, one might wonder why I have succumbed to an obvious inconsistency. In fact, the inconsistency is only in my use of the term 'reduction'. My true intention is only to argue that there are no *a priori* reasons for supposing that there could not be *some sort* of reduction of the mental to the physical. In defence of my infelicitous use of the term 'reduction', I attribute a similar fault to Davidson. Although his arguments against reduction are clearly meant to rule out both of the species of reduction defined above (I will speak of these two types of reduction together as consisting of *reduction proper*), it is also clear that Davidson intends his arguments to have stronger implications. To this effect he claims that there are "other forms of reduction" imaginable besides "direct elimination through [the] definition of psychological terms" in terms of concepts more like the concepts used in the physical sciences.(Davidson 1973a, p.246) Davidson has also claimed that "there is no important sense in which psychology can be reduced to the physical sciences"(Davidson 1973a, p.259), which is taken broadly to include biology and

physiology.(Davidson 1973a, p.253) Of equal significance as indicators that Davidson intends his arguments to have stronger consequences than merely that the mental is not *reducible proper* are claims (he has made) regarding the possible impact of detailed knowledge of the physics (or physiology) of the human brain on our knowledge of human psychology. In particular, Davidson has claimed, "If I am right, then, detailed knowledge of the physics or physiology of the brain, indeed of the whole of man, would not provide a shortcut to the kind of interpretation required for the application of sophisticated psychological concepts."(Davidson 1973a, p.258)

Davidson makes no serious attempt to make precise what lesser sorts of psycho-physical reduction he intends his arguments to rule out. Despite this lacuna, several of his claims about the relevance of the physical (construed broadly) to the psychological appear to be false. Furthermore, I believe that lesser forms that would be of interest could be achieved. I also believe that the arguments I have just accepted as precluding reduction proper do not rule out some lesser, yet still interesting, forms of psycho-physical reduction.[1]

§4.2 Lesser Forms of Reduction:

One such lesser form of reduction is the construction of what I will call *instantiation theories*. This notion is a generalisation of Robert Cummins' notion of an instantiation law. According to Cummins, instantiation laws are "statements specifying how a property [or state] is instantiated in a specified type of system."(Cummins 1983, p.7) In turn, I will call a set of such laws an *instantiation*

---

[1] I think that it is also clear that Davidson's conclusion that the mental is not reducible to the physical (as it appears in his early writings) is not meant to follow from either of what I have called the two strongest arguments against the possibility of psycho-physical reduction.

theory just in case the set of laws specifies for each of the states which correspond to the vocabulary of a selected theory, how each of the states is instantiated in a specified type of system. (An incomplete set of instantiation laws would be called a *partial instantiation theory*.) The degree to which such a theory would be of interest would obviously depend on the generality of the classes of systems to which the theory applied, and on the strictness and completeness of its laws.

Evidently instantiation theories of the mental are not ruled out by the multiple realisability of mental states. Yet the fact that the content of mental states can be determined by external relations may appear to be an impediment to the construction of such theories. The standard response to this apparent difficulty has been to deny the claim that the content of mental states is (or need be thought of as) determined by external relations. This response, is probably indefencible (even for one possessed of the view that semantic theory as a valid discipline ought to be brought into the service of science). Another approach to the apparent difficulty has been outlined by Jerry Fodor (1994). Simplified, and paraphrased in the terms of the construction of instantiation theories, Fodor provides reasons for recognising the *possibility* of discovering non-strict instantiation laws for mental states where exceptions to the instantiation laws are very rare. In effect, Fodor acknowledges that since the contents of mental states are determined by broad causal relations, it is possible that identical instantiating states, typed by their functional (or physiological) properties, could instantiate intentional states with different contents. Despite this conclusion, Fodor argues that it is at least possible, and, in fact, plausible that we could construct theories composed of non-strict instantiation laws for which exceptions would be uncommon.

If Fodor is right about the possibility of constructing such theories (and I think he is), then considerations of broad content show, at worst, that some of the instantiation laws which constitute an instantiation theory for the mental would not be exceptionless. This, however, does not amount to grounds for concluding that these non-strict instantiation laws would not support justified inferences to the token identity of particular mental and physiological states, and, similarly justified inferences from an agent's possession of a physiological state type to her possession of a mental state type. Moreover, despite the fact that the laws of an instantiation theory of the mental could counsel inferences to false conclusions, it does not follow that such laws could not support exceptionless inferences to an agent's possession of a particular *propositional attitude type*. For example, although we might, unaware of his bizarre origins, mistakenly attribute to a twin-earthling a 'fear' of going swimming in the in the icy H_2O, considerations of broad content do not in themselves rule out the possibility of our correctly attributing 'fear' to the twin-earthling solely on the basis of his physiological states. What considerations of broad content force us to recognise is just this: we can expect our instantiation laws to fail us on occasion, since we can expect the laws to be blind to differences in content in such cases where the object of a mental state possesses a 'double' which tends to produce identical causal impacts upon thinkers.

That Davidson thinks that it is impossible to construct instantiation theories for the mental is evident from statements to the effect that "our detailed understanding of the physical workings [of a subject], in itself, cannot force us to conclude that [the subject] is angry, or that he believes that Beethoven died in Vienna. In order to decide this, we would have to observe [his] macroscopic

movements, and decide how to interpret them".(Davidson 1973a, p.250) Davidson's claim indicates that he thinks that there cannot be instantiation laws for mental states, since the existence of an instantiation law implies that, given a system of a specific sort, it is sufficient for the system instantiating a specified state that it components are organised in a specified manner. Moreover, despite my acknowledgment in the preceding paragraph that some instantiation laws may fail (at least infrequently), it is far from obvious that there could not be instantiation laws for some mental states which would not be subject to exception. For example, it is feasible that laws for the instantiation of thoughts which do not concern objects external to an individual, such as thoughts about one's thoughts, would hold unconditionally.

§4.3 Davidson Against Instantiation Theories:

Having established that Davidson's argument for the irreducibility of the mental is also meant to preclude the discovery of instantiation theories of the mental (or even instantiation laws), I will now consider Davidson's argument. I proceed by considering the part of Davidson's argument about which there appears to be consensus.[2]

According to Davidson, the irreducibility of the mental to the physical follows from a divergence in the norms employed in mental and physical modes of explanation. As I wrote in section 1.2, Davidson thinks that the principles of psychological explanation differ fundamentally from other kinds of explanation in

---

[2] I will accept much of William Child's sympathetic interpretation of Davidson's somewhat cryptic argument, since I think that Child's work (1994) amounts to an accurate representation of the general structure of Davidson's reasoning.

virtue of the fact that psychological explanation is governed by a constitutive ideal of finding a rational order amidst the descriptions of mental events. Davidson concludes that since the mental and physical modes of explanation are governed by different constitutive norms, and because the constitutive norms of the mental cannot "be stated in a purely physical vocabulary" that the mental is not reducible to the physical (or the biological or the physiological).(Davidson 1973a, p.269)

The dictate that we find rational order amidst the descriptions of mental events is intended to imply that we should attempt to attribute to agents those mental states that they ought rationally to possess. On Davidson's estimation, the considerations that dictate what it is rational to believe and in what manner it is rational to act are holistic in the sense that in each case the factors relevant to applying the norms of rationality are potentially dependent on every feature of the case. In particular, Davidson claims, "There is no assigning beliefs to a person one by one ... for we make sense of particular beliefs only as they cohere with other beliefs".(Davidson 1970, p. 221) In short, Davidson thinks that determinations of what is rational are normative and holistic, and hence the criteria for attributing particular intentional states cannot be given. Finally, Davidson's conclusion is that since the application of mental concepts is guided by the dictate that we attribute those mental states which an agent ought rationally to possess, but since criteria cannot be given for what intentional states it is rational to possess, we cannot give criteria for the application of mental concepts.

While I believe that the preceding accurately reconstructs Davidson's reasoning, I also think that, as it stands, the reconstruction demands some

clarification. To this end I consider different ways of construing Davidson's reasoning. For ease of reference I will use the following analysis:

(1) We should attribute to agents the intentional states that they ought rationally to possess.

(2) Criteria cannot be given for what intentional states it is rational to possess.

(3) Since criteria cannot be given for what intentional states it is rational to possess, we cannot give physicalistic application conditions for what intentional states an agent ought to possess.

(C) We cannot give physicalistic criteria for the application of mental predicates.

I take it that this reconstruction is not uncharitable. Yet it seems clear that premiss 1 needs to be clarified. Indeed, without some instructions as to how we are supposed to interpret premiss 1, it is unclear whether the argument is valid.

I will say that a methodological dictate is *generally superseded* if and only if the application conditions of a predicate (to which the dictate is supposed to apply) have been fixed in such a way that the satisfaction of application conditions of the predicate do not logically (or at least nomologically) imply the satisfaction of the dictate. If the dictate (premiss 1) is the sort of methodological dictate which has the potential of being generally superseded by some other methodological dictate, then it appears that the argument is invalid, for if the demand can be generally superseded, then the uncodifiability of the norms of rationality do not carry over to

the uncodifiability of the application conditions for mental predicates.[3] Indeed, although it may be a dictate of the methodology of interpretation that we attribute to agents the intentional states which they ought rationally to possess, we need some grounds form thinking that this dictate could not be generally superseded.

We can distinguish at least two ways in which the application conditions of the concepts associated with a mode of explanation (for example, physical explanation, or mental explanation) could be uncodifiable. They are uncodifiable in one sense, if there could not be canon of methodology associated with the mode of explanation to which we can appeal in every case for the sake of making definitive arbitrating judgments between competing theories about the applicability of a concept (for example, the concept of mass) which both theories employ. This first manner of uncodifiability seems to characterise both the physical and mental modes of explanation. I do not take this sort of uncodifiability to rule out psycho-physical reduction, since it does not imply that we cannot, relative to a theory, codify the application conditions for the vocabulary of the theory. On the other hand, a second manner of uncodifiability does appear to preclude psycho-physical reduction (and the construction of instantiation theories). This second type of uncodifiability is simply the uncodifiability of the application conditions for the concepts central to the mode of explanation throughout the incarnations of the vocabulary as distinct theories.

---

[3] I will also say that the application conditions of a concept (or set of concepts) is uncodifiable if and only if criteria cannot be given for the application of the concept (or set of concepts).

Davidson obviously thinks that the application conditions of mental concepts are uncodifiable in the second sense. Moreover, he must adhere to a strong reading of premiss 1 to the effect that the methodological demand of attributing rationality cannot be generally superseded, since the general superseding the demand would permit a theory-relative codification of the application conditions of the mental vocabulary. That he thinks that the methodological demand cannot be generally superseded is evident from his claim that "when we use the concepts of belief, desire, and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase of the evolution of what must be an evolving theory."(Davidson 1970, p.222-3) Thus, while the preceding reconstruction of Davidson's reasoning is sound, it seems that it cannot be complete, since no part of the reconstruction explains why Davidson thinks that the methodological dictate of finding rationality cannot be generally superseded.

As a matter of fact, I think that it is an explicit form of behaviourism which serves as implicit support for the first premiss in Davidson's argument against the possibility of psycho-physical reduction. In the next two sections I will discuss the relationship between this Davidsonian behaviourism and the claim for the irreducibility of the mental.

§4.4 Behaviourism and Irreducibility:

Davidson repeatedly says with regard to radical interpretation that the evidential basis for interpretation consists in observations of what 'external circumstances' obtain when particular sentences are 'held true'.(Davidson 1974b, p.144 & 148)

Furthermore, Davidson holds that "a speaker holds a sentence true, because of what the sentence (in his own language) means, and because of what he believes. Knowing that he holds a sentence true, and knowing the meaning we can infer what he believes".(Davidson 1974b, p. 142) Indeed, holding a sentence to be true in relation to the external environment is, according to Davidson, accounted for by an agent's beliefs and what the sentence means in the agent's idiolect. In turn, judgments of what a person believes are generated on the assumption of a particular theory of truth for the person's language, and this theory is generated within the constraints of a behaviourist methodology. Finally, such theory-relative judgments about an agent's beliefs are taken by Davidson to be constitutive of what an agent's beliefs are. This fragment of Davidson's theory of *radical interpretation* is also in harmony with his claim that complete knowledge of the brain's functioning would provide no "short cut" to interpretation. Indeed, Davidson thinks that knowledge of the physiological states of an individual are only relevant to interpretation insofar as knowledge of an individual's physiology would permit us to determine the individual's behavioural dispositions. Knowledge of an individual's behavioural dispositions could in turn be included among the individual's actual behaviour (or exercised dispositions) as part of the sole evidential basis for interpretation.(Davidson 1973a, p.258)

It is useful to think of Davidson as adhering to the view that the *criterial* bases for the attribution of mental states are exclusively behaviouristic. This characterisation of Davidson's view is faulty, since Davidson does not think, as does someone such as Norman Malcolm or Gilbert Ryle, that there are *strict behavioural criteria* for the application of mental predicates. In fact, it is precisely Davidson's

disagreement with Malcolm and Ryle which motivates his claim for the irreducibility of the mental. Despite the admitted faultiness of the *preceding* characterisation of his view, it is clear that Davidson does think that the use of mental predicates is tied *exclusively* (and *essentially*) to behavioural *evidence*. Moreover, I believe it is this assumption which lies at the bottom of his belief in the irreducibility of the mental.

The line of reasoning which appears to lead Davidson to the conclusion that the mental is not reducible to the physical parallels the line of reasoning (I accepted as justified) for the conclusion that mental explanation has the distinctive characteristic of implicitly presupposing the rationality of the objects of interpretation. The discovery that led philosophers to the conclusion that interpretation presupposes the rationality of the subjects of interpretation was that there appeared to be a principled reason for why there could be no definitional reduction of the mental to the behavioural. It is as Davidson recognises, "we know too much about thought and behaviour to trust exact and universal statements linking them."(Davidson 1970, p.217) The well known problem is that in attempting to express behavioural application conditions for mental predicates we must invariably take recourse to non-behavioural mental concepts, for "no matter how we patch and fit the non-mental conditions, we always find the need for an additional condition (provided he *notices*, *understands*, etc.) that is mental in character."(Davidson 1970, p. 217) Accepting the conclusion that there could be no definitional reduction of the mental to the behavioural, it was concluded that an assumption of the rationality of subjects of interpretation had to serve as a basis for interpretation.

In addition to recognising the past and present behaviouristic basis of interpretation, Davidson has concluded that the evidential basis of interpretation is

exclusively behavioural. Given this conclusion, Davidson has made the further inference that the methodological dictate that we attribute to agents the intentional states they ought rationally to possess, cannot be generally superseded. Moreover, if we accept the claim that the methodological dictates of finding rationality cannot be generally superseded, along with the other premisses of my reconstruction of Davidson's reasoning, then his conclusion follows: we cannot give physicalistic criteria for the application of mental predicates.

Contrary to Davidson's conclusion, it should be pointed out that if we had an instantiation theory for the mental (which satisfied certain adequacy conditions), then the problem of discharging mental concepts from the conditions we use to rationalise exceptions to mental generalisations would be unproblematic. That is, if deny the assumption that the evidential basis of interpretation need be exclusively behavioural, then there is no *a priori* reason to suppose that we cannot have instantiation laws for the mental in physiological systems.

I will call an *instantiation* theory for the mental is *adequate* if and only if the auxiliary law is satisfied as a matter of nomological necessity for the systems for which the theory is for.[4] The idea is this: if we possess an *adequate instantiation theory for the mental*, then when an individual fails to behave in accord with his or her intentional states we will be able to show why the inefficacious states have failed to have their normal causal impact on the agent's behaviour. Typically, such an explanation would amount to pointing out a causal gap in the system in which the intentional states are instantiated. Such explanations would obviously be explicable

[4] The auxiliary law demands that any inconsistency that we attribute to an agent must be explicable.

in non-mental terms. Moreover, there is no good reason to suppose that it would be impossible to construct *an adequate instantiation theory* (for some intentional systems). For presumably any instantiation theory which identifies a system's representational states with objects whose formal properties are identical to their physical properties, the *auxiliary law* would be satisfied as a matter of nomological necessity.

In the next section I anticipate a possible response to my proposal for enlarging the evidential basis of interpretation. In doing so, I hope to forestall a possible confusion which might be thought to support the claim that the evidential basis of interpretation must be exclusively behavioural.

§4.4.1 A Strong Sense in which Internal States Could Be Relevant to Interpretation:

Someone might think that prior interpretations of an agent's behaviour are what set constraints on what physiological states we could identify as the agent's mental states. Such an idea seems to be the only explanation of why (a long time ago) Dennett denied the possibility that behavioural duplicates could possess different intentional states.(Dennett 1978) This conclusion *seems* to follow from an assumption we ought to accept: that the function of intentional psychology is to predict and explain *behaviour*. Thus, behaviour provides constraints to which our physiological discoveries must conform, if we wish to identify mental states with physiological states.

Despite the truth of the assumption, the conclusion that prior interpretations of behaviour set the constraints does not follow. The reason is generally accepted:

any body of (non-intentionally described) behavioural evidence will support many different sets of attributions of mental states. Thus, while behaviour does provide the constraints, no particular set of mental states is implied by such a restriction. It may be, then, that through componential analysis of the bases of behaviour, physiological research could aid us in choosing between interpretations of some individual that are equally confirmed by behavioural evidence. It may also be that physiological discoveries will point irresistibly in the direction of interpretations of subjects that would never have seemed plausible (for methodological reasons) or even conceived of on the basis of behavioural evidence alone. The fact that physiological evidence could assist us in choosing among interpretations of behaviour (within an acceptable range) implies that there could be principled reasons for attributing to behavioural duplicates different sets of intentional states. In fact, even pairs individuals whose behavioural dispositions were identical could possess different intentional states, since there is no good reason for denying that an individual could possess an intentional state which could not be acted upon. That is, we can imagine an organism in a type of functional state which has a disposition to influence the behaviour of the organism, even though the particular state could never actually do so, because of the other functional states of the organism.

The degree to which an interpretation of an individual derived from a componential analysis *via* physiological research could diverge from a prior interpretation derived exclusively from behavioural evidence may be constrained to some degree. But there is no reason to think that relatively to a particular class or species of agents the application conditions for a mental predicate could not be

identical to the application conditions for a physiological predicate (or a predicate of some other functional type). (Consider, for example, the thinking machine that could be built for the sake of demonstrating this fact.)

To the degree that Davidson would deny the possibility of such identifications, it seems that his non-reductionism (which I take to preclude instantiation theories for the mental) is supported by an illicit assumption about meaning. This assumption is that a change in the application conditions for mental concepts, which permits their application on the basis of non-behaviourally described facts, would entail a change in the meaning of the mental predicates to such a degree that we could no longer describe our attributions as attributions of mental states.

§4.5 Can there be Reduction (or Instantiation Laws) without Changing the Subject?

Davidson's claim that the reduction of the mental to the physical would change the subject is rooted in his views concerning the requirement for the preservation of a particular kind of discourse and a corresponding discourse-relative conception of humanity. This motivation is made evident by several things that he says in his paper *Mental Events*. Davidson argues that existence of strict laws linking the mental and the physical could only be granted on pain of "changing the subject", and thereby "deciding not to accept the criterion of the mental in terms of the vocabulary of the propositional attitudes."(Davidson 1970, p.216) Similarly, Davidson has argued that the maintenance of a degree of "nomological slack between the mental and the physical is essential as long as we conceive of man as

a rational animal."(Davidson 1970, p. 223) It is fairly clear from statements such as this that Davidson thinks that the maintenance of a mental vocabulary in a particular state is necessary for the maintenance of a certain self-conception. One might concede this claim taken broadly, though it is unclear (at least to me) what Davidson thinks would be sacrificed by the reduction, broadly construed, of the mental to the physical, again broadly construed. The maintenance of a vocabulary which preserves certain reflexive descriptive functions is necessary for the sake of conducting particular moral and epistemological practices. But the question remains: what sort of features must a vocabulary maintain for the sake of fulfilling such moral and epistemological practices?

It is difficult to give an accurate outline of the function of mental explanation (and its associated concepts) which is both brief and accurate. But I will attempt to give a sketch that gets at the *practical* difference between this and other modes of explanation. The first point to be made is that the mental vocabulary is predictive and explanatory in much the same way as other truth-conditional vocabularies. When we deploy our mental vocabulary we assume a particular explanatory and predictive stance distinguished from other stances by the manner in which it is used to explain the behaviour of objects (that is, by the attribution of intentional states). But the mental vocabulary is also essential for performing certain tasks. In particular, this vocabulary is useful to the degree that it permits self-description for the purposes of applying knowledge collected over centuries to the process of reflection upon what course of action to pursue, or what to believe. Here the correspondence between thought and language permits the application of accumulated knowledge. The use of the mental vocabulary also seems to be

essential to the application of accumulated knowledge to the practice of manipulating the behaviour of agents (besides oneself). It is chiefly because of its capacity to permit us to perform such functions that mental explanation has advantages over explanation in the physical mode.

With regard to Davidson's claim that reduction would 'change the subject', we can reply that the only changes that need be brought about (or could be, without changing the subject) for the sake of the construction of instantiation theories for the mental would be an adjustment of the application conditions for mental predicates, and this *could* be achieved without changing the forms of conceptual relations which obtain between intentional state types (i.e., beliefs, desires, intentions to act, etc.). The point is this: if our psychological theories preserve the conceptual, and dispositional relation patterns between the intentional state types, then there is no reason to think that instantiation theories reflecting a shift in the application conditions of the mental predicates would fundamentally change the practical concerns which could be serviced by such a theory. Desires, beliefs, and so on would still be related as they always had been, and the facilitation of our concern for the manipulation of beliefs, desires and intentions to act would remain intact.

§4.5.1 A Minimalist Account of the Nature of the Folk Theory:

Contrary to what might have been suggested by my arguments in chapter 1, folk psychology should probably not be thought of as a single theory (though for ease of expression I will continue to speak as though it were). Rather the practices which fall under the extension of this expression can be viewed as guided by a family of theories, sufficiently similar that consensus generally can be reached about the

applicability of particular mental predicates in particular instances. What is common to all these theories can be captured by a *form* common to all the theories that guide our interpretive practices. This from is embodied in the core laws of intentional psychology, and is preeminent over *application conditions* (the other component of any theory) as the *essential* feature of the mental explanation. Though as is the case with most modes of explanation, and certainly in the case of intentional psychology, some conceptual features are essential to the mode of explanation beyond the relations expressed by its core laws. (In the case of intentional psychology, features of our concept of intentional state demand that an agent's possession of such a state entails his possession of an appropriate contentful state.)

The idea that the essence of intentional psychology is constituted by the three core laws (along with the auxiliary law and the coherence principle) is similar to Davidson's conception of intentional psychology. I have argued, however, that Davidson had it wrong: instantiation theories for mental states *could* be constructed without changing the subject. But it may be asked in turn: what sort of considerations support the claim that the core laws of intentional psychology, in conjunction with the auxiliary law, are definitive of the subject of intentional psychology? What follows is a roundabout answer.

Rejection of Davidson's claim for the irreduciblity of the mental (construed broadly) is warranted by considerations of minimalism. Davidson's view on reduction (and similarly his opposition to the possibility of instantiation theories) seems ultimately to depend on a view about conceptual change, a view that manifests itself in the demand that we tie the identity of mental concepts to an exclusively behaviouristic methodology for the attribution of mental states. This view

is no less extravagant than Norman Malcolm's analytic behaviourism. Malcolm is known to have criticised psychologists, who claimed to study dreaming by studying such phenomena as REM sleep, for inventing and associating a new concept with the term 'dreaming'.(Malcolm 1962) Malcolm thereby denied that these psychologists were studying dreaming.

In looking to discover the essential features of our mental vocabulary we are looking for something like necessary and sufficient conditions.[5] It is on this assumption that Davidson's characterisation of what is essential to our mental vocabulary is rejected. We could generate instantiation theories for the mental, and thereby refer to the physiological states identified by such theories as the mental states of the instantiating agent, and the subject matter would still be the propositional attitudes of persons. Thus, Davidson's proposed necessary condition on instances of mental explanation must be peeled away. What is left may be simply the present view.

Let us now consider the degree to which the form of intentional psychology constrains the application of mental predicates. I would like to make some precise claims about the relation between the mental and physical vocabularies, and in particular about the type of constraint on the construction of physicalistic instantiation theories for mental states that must be acknowledged. The constraint is just this: instantiation laws and empirical generalisations featuring the relation of mental predicates do not implicitly define the concepts that correspond to the names of the mental predicate types. Rather these concepts are defined (at least in part) implicitly by the core laws that constitute the form of psychological theories. In other

---

[5] Though I doubt we could succeed in articulating such conditions.

words, although the nominal concepts corresponding to intentional state types can be equated with strict (or at least fairly strict) domain-specific (or species-specific) physiological application conditions, the real essence of the concept associated with various intentional state types is not bound to the nominal application conditions for the state types. Each alteration of the application conditions for a mental predicate is beholden to the formal principles of intentional psychology, and thus, the mental predicates maintain their allegiance to these principles. But what does this allegiance amount to? In other words, to what degree and in what manner does the form of psychological theories constrain the application conditions for mental predicates?

§4.5.2 Reduction and the Norms of Interpretation:

No precise claim can be defended regarding the degree to which the form of psychological theories, and it associated methodology constrain the physicalistic application conditions for mental predicates. As we have seen, it may be that as a matter of nomological necessity a instantiation theory for the mental would satisfy the auxiliary law. On the other hand, the dictates of the coherence principle constitute merely one dictate among many which may have an impact on the psychological theories we construct (with this principle being peculiar to intentional psychology). While a choice of application conditions for mental predicates that maximises coherence will generally tend to increase the interpretations projectability, the urge to eliminate all exceptions is unreasonable, if we want our theory to apply to human behaviour. In fact, there is no good reason to suppose that the coherence principle could not be generally superseded.

In chapter two I distinguished *strategic* and *methodological* reasons for assuming the rationality of subjects of interpretation. Strategic grounds for assuming the rationality of subjects of interpretation derive from the assumptions about how sets of intentional states tend to be tested by the world, and for this reason such grounds are defeasible, since such assumptions concern the history of an subject, and may turn out to be false. In principle, we could discover everything about the history of an agent, and would need no heuristics implicating generalisations about his *probable* history. On the other hand, the methodological consideration that enjoins us to maximise coherence is independent of any empirical assumptions about the history of the individuals we interpret. The coherence principle is justified by the fact that in adhering to it we ensure that sets of intentional states we attribute will serve as a sound basis for *making predictions*. Though this methodological assumption is not based on falsifiable empirical assumptions, there are still good reasons to think that the application of the principle could be generally superseded. The generation of a theory which regularly permits attributions of incoherent belief sets (especially in the vicinity of an insane asylum), could be quite valuable, especially if it comes with a body of auxiliary hypotheses for rationalising exceptions to the core laws, and, again, especially if the instances of application for the auxiliary hypotheses are readily determinable. Precisely these conditions would be satisfied by the construction of an instantiation theory for mental states. Thus, the construction of such a theory would inevitably loosen the grip of the coherence principle. If an instantiation theory for mental states is derived via a successful componential analysis of the brain's realisation of human cognitive behaviour corresponding to the conceptual categories of intentional psychology, the

methodological purpose of the characteristic norm of interpretation, *the coherence principle*, would be outpaced and the dictate justifiably superseded.

There is, it must be admitted, a second way in which the projection of rationality may enter into the process of interpretation, and this deserves a brief discussion. It is generally accepted among philosophers that norms enter into interpretation as an indispensable basis for resolving indeterminacies of reference. Such norms play a role in fixing the distal referents of thoughts at places along the causal chains between agents and their environments. Such norms are probably ineliminable as a basis for the assignment of content.

Nevertheless I do not think that the role of such norms is an impediment to the construction of instantiation theories. The reason was touched on in section 4.2. In that section, I argued that considerations of broad content force us to recognise that we can expect the laws of any instantiation theory we construct to fail us on occasion, since we can expect laws of any instantiation theory to be blind to differences in content in such cases where the object of a mental state possesses a 'double' tending to produce identical causal impacts upon thinkers. Accepting that the impact of reference-fixing norms is not more extensive requires us to recognise the limited scope of such norms as a basis for settling questions of reference.

Reference-fixing norms come into play when the physical facts of a situation leave indeterminate a question concerning the reference of an agent's thought. However, while such norms may come into play in every case, I think that the amount of 'play' with regard to where such norms may invite us to fix the reference of a thought is relatively small. The point is that the scope of the referential indeterminacies that may be resolved by appeal to considerations of rationality are

typically small. This is because physical factors predominate in fixing the distal referents of an agent's thoughts. An analogy will illustrate the point. It is obvious that we often use heuristics (sometimes involving the projection of rationality) in order to decide what an individual is looking at. However, we can also get a good idea about what some one is looking at, if we know something about the way eyes work, and particularly about the way in which the eyes of organisms with binocular vision co-ordinate for the purpose of attending to objects. It does not appear that norms of rationality are applied in such ways of reasoning about what an individual is attending to. Similarly, it may be that we could determine a fairly exact approximation of the distal referents of a physiologically instantiated mental state by discovering functional similarities between the state and other states of the agent which have dissimilar causal relations to his environment. For example, there may be instances where we would take the systemic identification of information states (as revealed by functional similarities) which were engendered by disparate proximal relations to the agent's environment as indicating an identification of the states with respect to their distal referents. Given such states, we may be able to perform a sort of *triangulation* along the causal antecedents of the mental states to determine a common causal antecedent, and, in effect, the distal referent of such states.

In defence of my brief discussion of what is obviously a very complicated issue, one should remember that the claim which I am attempting to establish is modest. I am merely attempting to establish *the possibility* of constructing physiological instantiation theories for the mental.

§4.6 The Prospects for the Construction of Instantiation Theories:

Some clarification of my position is called for. While I have argued earnestly for the possibility of the domain-specific identification of mental states with physiological states (through the construction of instantiation theories), I have no conviction about the degree to which such a possibility will be actualised (as measured by applicability of a particular instantiation theory to a large class of objects, or by strictness and completeness). And I do not believe that our mental vocabulary needs to be vindicated by the construction of physicalistic instantiation theories. On the other hand, the type of naive realism to which I subscribe (which compels me to think that mental states could be identified with physical states) follows from two beliefs: one, about the nature of semantics and the relation of semantics to practice; and, another, about how human intuitions tend to be manifested in semantic practice. The best way to explain this woolly view is by commenting on the way I (perhaps idiosyncratically) understand a passage from Wittgenstein's *Blue Book*. In this book, after many pages of discussing how it is that certain forms of words do not have sense, Wittgenstein mentions a type of confrontation which might occur between a philosopher and a scientist. The confrontation occurs when the philosopher asserts that the form of words 'unconscious pain' has no sense, and a scientist retorts that not only does the expression *have* a meaning, but he has proven that people *have* unconscious pains. Wittgenstein drew his own moral from the story. My own feeling has always been that the scientist should have been commended for his discovery, and the philosopher sent to find a project less disruptive to the progress of science. My attitude towards Davidson's behaviourism about mental state attribution is much the same. If scientists find structures in the

brain suitable for the identification with mental states (and in particular propositional attitudes), they will make the identification and assert the identity as a matter of fact. If such identifications *are* made Davidson should stand aside.

I take it that I have shown that it is possible that such identifications *could* occur. Such identifications may occur by the sheer force of the will of cognitive scientists attempting to develop a componential analysis of the brain's functions corresponding to the categories of intentional psychology. I also think that the sort of matter-of-fact realism generated by such identifications would be and should be accepted. Of course, such a matter-of-fact realism does not preclude the revisability of theories, and psychological theories are no exception. This remark leads naturally to some final observations about the prospects of folk theory.

§4.6.1 The Prospects for Elimination:

In the introduction to this essay I claimed that there are two good reasons for defending the claim that our folk practice of attributing mental states is theory-driven. The first is that it is true. The second is that it permits us to recognise the genuine possibility that the theory in which the practice is grounded might be false.

In itself, my conclusion that the folk practice is not guided by a single theory makes little difference to the present issue, since the conclusions offered by eliminativists are general and are intended to entail the elimination of intentional psychology. With respect to intentional psychology, generally eliminativists are

committed to the claim that there are no such things as intentional states (or at least that we humans do not have them).[6]

At least two species of argument have been deployed for the conclusion that the elimination of our folk theory is imminent. The more prominent (recently used by Stich (1996)) is that intentional psychology presupposes that cognition has a particular sort of functional structure and that empirical evidence is beginning to disclose and will eventually show that the brain simply does not possess a componential structure that could be construed as realising the functional structure presupposed by intentional psychology. The view is that alternate models of cognition (and possibly connectionist models) will prevail and that the illegitimacy of intentional psychology will be demonstrated by this outcome. Andy Clark refers to this position (which, he admits, may not be held by Jerry Fodor) as Super-Fodorian Realism, since adherents of the position believe that if such a view of the mind as is held by Fodor turns out to be false, then so does the folk theory.(Clark 1993) Needless to say I agree with Clark, and also John Heil (1991), both of whom have argued that so-called Super-Fodorian Realists come up with grounds for the elimination of intentional psychology only because they make unjustified claims about the sorts of componential structures to which intentional psychology is committed.

The second species of argument for eliminativism comes from Paul Churchland.(1981) Churchland characterises folk psychology as a stagnant research project which will be eliminated in virtue of its inability to give an adequate

---

[6] Some Philosophers, calling themselves eliminativists, have argued for the weaker conclusion that the categories of intentional psychology will not feature in the ontology of a mature cognitive science. I will not consider this sort of eliminativism.

explanation of certain aspects of human behaviour, and in virtue of its inability to compete with blossoming models of cognition which are being developed by people such as connectionists. Setting aside the issue of whether the folk theory deserves Churchland's scorn, he has overlooked an important question. Will the new theories of cognition actually be competitors of the folk theory. This much is clear: the new models of cognition may subsume some of the explanatory and predictive domain which is now attended to by intentional psychology. But it is unclear that the new theories could attend to all of the functions serviced by intentional psychology. Thus, it is unclear why such theories cannot coexist. And here I return to a point similar to one raised by Davidson. Recall that Davidson argues that existence of strict laws linking the mental and the physical could only be granted on pain of "changing the subject", and thereby "deciding not to accept the criterion of the mental in terms of the vocabulary of the propositional attitudes."(Davidson 1970, p.216) His warnings are meant to expose the fact that in the case of intentional psychology reduction entails elimination. I have tried to show that no such entailment holds. I now wish to emphasise that elimination entails "deciding not to accept the criterion of the mental in terms of the vocabulary of the propositional attitudes." Moreover, the considerations Churchland raises do not license the elimination of intentional psychology, if the proposed succeeding cognitive science does not subsume all of the important explanatory functions serviced by intentional psychology.

The problem with supposing that intentional psychology will be eliminated by a succeeding science of cognition on the basis of its explanatory inadequacy is that of supposing that any other mode of cognitive explanation could fulfill the

explanatory functions of intentional psychology. In the case of criminal law we have an example of a practice where no theory of human cognition but the folk theory will do the job. Indeed, "Crimes are defined by statutory or judicially created (common law) criteria known as the elements of the crime. The definition of all crimes include a prohibited act (or omission, if there is a legal duty to act) and, with few exception, an accompanying mental state."(Morse 1992, p.207) But if a 'new' science of cognition will not fulfill the functions of intentional psychology, then the new theory is not in some significant respects its successor. In that case, the considerations which Churchland raises as grounds for the elimination of intentional psychology are uncompelling.

A final response can be given in defence of the considerations raised by Churchland: even if a new science of cognition is not capable of fulfilling all of the former functions of intentional psychology, there is no reason why we cannot accept intentional psychology as disconfirmed, yet maintain its use in a manner like that of Newtonian mechanics. The idea here is that the theory is shown to be physiologically disreputable and is hence we conclude that humans do not possess intentional states. Despite the repudiation of intentional psychology, its application is maintained in an instrumentalist form, thereby continuing to satisfy it former functions. In the next section I discuss the possibility that the state of intentional psychology is already aptly described as instrumentalist. I reach the conclusion that a realist commitment has always been implicit in intentional psychology, and that it is essential to intentional psychology that its application presupposes that an agent's being in an intentional state implies his being in an appropriate contentful state, one

that is a causally efficacious and tending to determine the agent's behaviour in ways appropriate to its content.

§4.6.2 Instrumentalism with regard to Intentional States:

Dennett has on occasion referred to his view of intentional state attribution as instrumentalist. What he really means by this is that the truth of claims about the intentional states possessed by an agent depend on no picture of the physiological states that realise the causal bases of the agent's behaviour. One is entitled to ask: for Dennett, what are the bases for evaluating the truth value of attributions of intentional states? Dennett's view about the assignment of truth values to such attributions is, roughly, that the true attributions are those that are explanatory and serve as the basis for accurate predictions. Similarly, intentional agents are those objects whose behaviour can be adequately explained and accurately predicted by the attribution of intentional states. Obviously something is amiss with this account of the semantics of mental state attribution.

Our reaction to an example proffered by Dennett himself is symptomatic of the problem that Dennett himself acknowledges. In his *Reply to Professor Stich* (1980), Dennett admits that "the claim that human beings are genuine believers and desirers" is not "in principle" invulnerable to scientific disconfirmation, "for in a science-fiction mood we can imagine startling discoveries (e.g., some "people" are organic puppets remotely controlled by Martians) that would upset any particular home truths about believers and moral agenthood you like, and - more importantly - a partial erosion of our self-image as rational, self-controlled agents due to discoveries about our cognitive imperfections is not ruled out."(Dennett 1980, p.73)

It seems that what Dennett is admitting here is inconsistent with the claim that the truth conditions of mental ascription rest on no picture of the physiological states which realise the causal bases of an agent's behaviour. In fact, it is not clear that this is so. If one recognises that Dennett is justifiably characterised as an extreme minimalist about intentional psychology, then one would expect that he would tolerate the disconfirmation of many (and possibly all) of the presuppositions which are generally thought to constrain the application of mental concepts. If we forsake such presuppositions, Dennett acknowledges that we would have given up our conception of ourselves as "genuine" believers. However, Dennett does not acknowledge that such changes in our self-conception would alter the semantics of mental state attribution.

Contrary to Dennett, and in spite of my own minimalist leanings, some factors *essential* to the semantics of mental state attribution are not embodied in the core laws of intentional psychology. These essential factors are distinct from the sorts of "home truths" which Dennett acknowledges as eliminable. Like Dennett, I do not take a self image of ourselves as rational self-controlled agents to be *essential* to our regard for ourselves as the possessors of intentional states. But other features *are* essential to our conception of what it is to possess an intentional state.

One is that the subject has at one time stood in appropriate causal relations to the kinds of objects that serve as the contents of the subject's beliefs. Similarly, that the provision for explanation and prediction is not sufficient for the correct attribution of intentional states is evident from the fact that we are ordinarily unwilling to attribute to an agent a belief about a particular physical object when no causal *rapport* has existed between the object and the agent. The reason we behave this

way is known by anyone who understands the nature of intentional psychology. Though there are unusual cases, such as when one names an object using a definite description, in general, in order to have a thought about a particular physical object, the object must have exercised a causal impact upon the agent's cognitive apparatus through perceptual acquaintance or through second had reports. This impact must also have some residual influence upon the agent in the form of a contentful state, which in virtue of its content, has dispositions to cause the agent to perform actions appropriate to the state's content.

It is consistent with the observations just made that the truth-conditions for possession of an intentional state depend on no *particular* picture of the cognitive processes which realise the behaviour of the agent who possesses the state. Contrary to Dennett, some very general conditions concerning the relation of a believer to his objects of belief and concerning the causal structure of the physiological states which realise the subject's behaviour must be satisfied. Contrary to eliminativists, the way in which such conditions may be satisfied is very much open, and is surely not bound to a Fodorian picture of the mind. Providing a characterisation of what conditions would need to be satisfied is beyond the scope of the present essay. However, I suspect that such questions cannot be useful answered prior to our achieving a deeper understanding of human cognition.

Summary:

I Began Chapter one with a defence of the theory-theory. To this end, I first tried to demonstrate the faultiness of intuitions to the contrary. Second, I cited experimental evidence in favor of the conclusion that our folk practice is knowledge-driven. Finally, I tried to show that the our folk practice is implicitly constrained by three principles (the core laws of intentional psychology) which appear to provide for the systematic integration of our knowledge of the behaviour of sophisticated organisms. These core laws appear to serve as a basis for systematising our knowledge of human behaviour (non-intentionally described) by implicitly defining concepts which permit representations of the fine-grained patterns of non-intentionally described human behaviour in terms of coarse-grained intentionally described patterns of behaviour.

If our practice is dependent on the application of the core laws and the concepts they implicitly define, then folk psychology possesses the important functional features characteristic of paradigmatic scientific theories. Moreover, as the best explanation of the of the three implicit laws, I offered the conclusion that something like the these laws are *tacitly represented* in human brains.

I acknowledged that it is difficult to demonstrate that a systematicity is implicit in mental explanation. However, considerations of the apparent normative and holistic bases of interpretation lends credibility to the claim. In addition to this, the normative and holistic character of intentional explanation implicated in the core laws of intentional psychology provides an explanation of how mental states are attributed despite the absence of strict criteria for the application of mental concepts.

In the later stages of chapter one, an explanation was provided of how the normative character of mental explanation is manifested in the methodology of interpretation. Two methodological dictates were endorsed. The first, *the auxiliary law*, demands that any inconsistency that we attribute to an agent must be explicable. The second, *the coherence principle*, counsels that we maximise the coherence of the sets of intentional states we attribute to agents. The intuitive plausibility and the apparent widespread acceptance of the auxiliary law supports the claim that normative relations represented in the core laws of intentional psychology generally constrain our use of mental concepts.

The function of the coherence principle is to ensure the projectability of the sets of intentional states we attribute. The way the coherence principle is applied also provides a basis for understanding the way in which the normative bases of interpretation work against the empirical bases of interpretation to the betterment of interpretations. I have already supposed that mental explanation is underpinned by an implicit understanding of the normative relations between state types. But despite the presumption that these relations are upheld in the course of interpretation, mental explanation permits the integration of empirical data as a basis for recognising the types of cases where one should make an exception to the core laws of intentional psychology. Thus, while the presumption of maximising the coherence of the sets of intentional states we attribute prevents the attributions of sets of intentional states which will not support predictions of behaviour, knowledge of the sorts of conditions under which the core laws are prone to exception constrains the application of the coherence principle. The partial suppression the

principle allows the application of knowledge about the regularity of patterns of intentionally and non-intentionally described behaviour to influence interpretation.

Aside from the well groundedness of the second dictate as a constraint on the attribution of unprojectable interpretations, the second dictate makes sense of the apparent normative basis of interpretation without leaving any difficulties in explaining our ability to attribute irrationality. The principle can be justifiably restrained on the basis of considerations derived from empirical data regarding the regularity of particular patterns of behaviour.

By formalising several concepts of *coherence*, and by schematising a method of interpretation which applies these formalised concepts, chapters two and three are intended to illustrate how the core laws and the two characteristic dictates of the methodology of intentional psychology might be applied in the practice of interpretation. The focus of the schematisation is on illustrating a general approach to rationalising exceptions to the core laws. I take it that the general approach is in accord with the most intuitive but informal way of understanding exceptions to optimally rational behaviour: we postulate limitations in cognitive ability.

Chapter three was devoted to answering several technical questions concerning the application of paraconsistent logic to the attribution of intentional states. The motivating supposition was that after attributing a coherence measure as a constraint on the sets of intentional states we will attribute to an agent, and after attributing a set of basic explanatory intentional states, we should permit ourselves to attribute intentional states as consequences of the basic explanatory intentional states. Such additional states would in turn establish possibilities for the

prediction of the behaviour of a subject which do not require the augmentation of the agent's basic explanatory intentional states.

Several desiderata of a logic for the proposed application were described. First, such a logic must be sound and complete with respect to a consequence relation which preserves a property, ψ, within closure, where ψ is a selected measure of the coherence of a set of basic explanatory states. Second, there are methodological grounds for employing the strongest consequence relation available. The first of these two desiderata is obviously satisfied inasmuch as it was shown that a well known consequence relation, *forcing*, preserves many properties within closure including a measure of coherence, *dilution*, which seems particularly apt for the proposed application. The second of the two desiderata is satisfied insofar as a lower limit n was discovered such that n-forcing does, but (n-1)-forcing does not, preserve dilution.

In chapter four, I tried to mediate several claims to the effect that the seemingly unique normative character of mental explanation implies the impossibility of psycho-physical reduction. I began by acknowledging that considerations of multiple realisablity and externalism probably rule out the possibility of *reduction proper*. Despite this admission, I argued that there are other interesting sorts of reduction. In particular, I argued for the possibility of constructing instantiation theories for mental states. In opposition to my proposal, it is clear that *it is* an intended implication of some the arguments for the irreducibility of the mental that even such modest forms of reduction are impossible.

In defence of my claim for the possibility of constructing instantiation theories, Davidson's version of what I take to be the principle argument against psycho-

physical reduction is considered. I reach the conclusion that this argument is dependent on the assumption that the methodological impact of the characteristic principles of the methodology of interpretation (that we project rationality into the sets of intentional states we attribute) cannot be *generally superseded.*

I considered the two characteristic principles of the methodology of interpretation in turn. With regard to *the auxiliary law,* I argued that there is no good reason to suppose that we could not construct *adequate instantiation theories* for which the auxiliary law would be satisfied as a matter of nomological necessity. On the other hand, I argued that there is no good reason to think that *the coherence principle* could not with some justification be *generally superseded.* To make this point the methodological purpose of the characteristic norm of interpretation was reconsidered. In chapter two, I paid particular attention to distinguishing *strategic* from *methodological* grounds for projecting rationality into the sets of intentional states we attribute; that distraction serves at this point to circumscribe the scope of the demand that we project rationality into the subjects of interpretation. I conclude that the purpose of the coherence principle is to ensure the projectability of interpretations as a basis for making predictions of future behaviour. Since the construction of an instantiation theory for the mental would ensure the projectability of our physiologically grounded attributions of intentional states, I conclude that the coherence principle can justifiably be generally superseded.

To conclude chapter four, several consequences of my nearly minimalist characterisation of intentional psychology are drawn. My characterisation of intentional psychology is not radically minimalist, since I suppose that it is essential to intentional psychology that its application presupposes that an agent's

possession of an intentional state implies his possession of an appropriate contentful state. This fact is offered as a consideration against construing intentional psychology as an instrumentalist mode of explanation. In turn, given this essential property, it is conceivable that human beings might not have intentional states. Despite this acknowledgment, the characterisation of my view as nearly minimal is intended to emphasise the potential resilience of intentional psychology (as a mode of explanation) in virtue of its relative lack of ontological commitments. This resilience in conjunction with the legal and moral functions which mental explanation is uniquely capable of performing are offered as grounds against the likelihood of its elimination.

Appendix A:

$|\Sigma|$ = the number of elements in Σ.

$\wp(\Sigma)$ = the set of all subsets of Σ.

$\lceil x \rceil$ = x rounded up to the nearest natural number.

$\lfloor x \rfloor$ = x rounded down to the nearest natural number.

Bin(x) = the binary representation of a natural number x.

$at(\Sigma)$ = the number of distinct truth functional atoms which appear in Σ.

Appendix B:

Theorem: $\forall\Sigma: C(\Sigma) \leq m \Rightarrow \mathcal{A}(\Sigma) \geq \lfloor \log_2(1/m) \rfloor$.

Proof:

Let $B(\Sigma) = \{ a \subseteq \Sigma \mid a \vdash_{PL} \perp \}$.

Then, obviously, $\forall\Sigma: C(\Sigma) \leq m \Rightarrow |B(\Sigma)| \leq (2^{|\Omega|} \times m)$.

Moreover, inconsistency is monotonic,

that is, $\forall\Sigma: \Sigma \vdash_{PL} \perp \Rightarrow (\forall\Sigma': \Sigma \subseteq \Sigma' \Rightarrow \Sigma' \vdash_{PL} \perp)$.

Thus, if a set, Σ, possesses an inconsistent subset, a, of a certain size, n (which is implied by its having a dilution of n), then it must possess no less than particular number of inconsistent subsets, equal to the number of supersets of a in $\wp(\Sigma)$. Formally, the reasoning is as follows:

$\forall\Sigma: \mathcal{A}(\Sigma) = n \Rightarrow \exists a \subseteq \Sigma: |a| = n \ \& \ a \vdash_{PL} \perp$.

Moreover, $\forall b \in \{ b \subseteq \Sigma \mid a \subseteq b \}: b \vdash_{PL} \perp$.

But $|\{ b \subseteq \Sigma \mid a \subseteq b \}| = 2^{|\Sigma - a|} = 2^{|\Sigma| - n}$.

Thus, finally, $\forall\Sigma: \mathcal{A}(\Sigma) = n \Rightarrow |B(\Sigma)| \geq 2^{|\Sigma| - n}$.

Similarly, $\forall\Sigma: \mathcal{A}(\Sigma) < n \Rightarrow |B(\Sigma)| > 2^{|\Sigma| - n}$.

But since $\forall\Sigma: \mathcal{A}(\Sigma) < n \Rightarrow |B(\Sigma)| > 2^{|\Sigma| - n}$,

by contraposition $\forall\Sigma: |B(\Sigma)| \leq 2^{|\Sigma| - n} \Rightarrow \mathcal{A}(\Sigma) \geq n$.

It now becomes obvious that knowing a set's corruption will allow us to compute a floor for its dilution, for the role played by the size of the set can be factored out. Indeed, $(m \times 2^{|\Sigma|}) = (2^{|\Sigma|-n})$, when $n = \log_2(1/m)$.

Moreover, since $\forall\Sigma: C(\Sigma) \leq m \Rightarrow |B(\Sigma)| \leq (2^{|\Sigma|} \times m)$,
it follows that $\forall\Sigma: C(\Sigma) \leq m \Rightarrow |B(\Sigma)| \leq (2^{|\Sigma| - \log_2(1/m)})$.

But since $\forall\Sigma: |B(\Sigma)| \leq 2^{|\Sigma|-n} \Rightarrow \alpha(\Sigma) \geq n$,
it follows that $\forall\Sigma: |B(\Sigma)| \leq (2^{|\Sigma| - \log_2(1/m)}) \Rightarrow \alpha(\Sigma) \geq \log_2(1/m)$.

It follows that: $\forall\Sigma: C(\Sigma) \leq m \Rightarrow \alpha(\Sigma) \geq \lfloor \log_2(1/m) \rfloor$.

Appendix C:

Pascal's Triangle:

							1											
						1		1										
					1		2		1									
				1		3		3		1								
			1		4		6		4		1							
		1		5		10		10		5		1						
1			6		15		20		15		6		1					
				:		:		:		:								
				:		:		:		:								

The significant feature of this structure is that the value of each position (save the value of the positions of the top row) is equal to the sum of the two most adjacent numbers on the above row. Sequences of numbers which instantiate Pascal's Triangle tend to arise from time to time. For example, the following table which displays the composition of powersets is such an instance.

The Composition of Powersets:

The following table expresses the cardinality of the sets of subsets of the powerset of the set, according to their cardinality. For example, the powerset of a set whose cardinality is 8, will possess 56 subsets whose cardinality is 3.

$\lvert\Sigma\rvert$ =	1	2	3	4	5	6	7	8	9
$\lvert\wp(\Sigma)\rvert$ =	2	4	8	16	32	64	128	256	512
$\lvert\{ b \mid b \in \wp(\Sigma) \}\rvert$, where $\lvert b \rvert = m$.									
m=1	1*	2*	3*	4*	5*	6*	7*	8*	9*
m=2		1	3	6	10	15	21	28	36
m=3			1	4	10	20	35	56	84
m=4				1	5	15	35	70	126
m=5					1	6	21	56	126
m=6						1	7	28	84
m=7							1	8	36
m=8								1	9
m=9									1

*-the empty set is not counted.

The following table represents the respective ceiling for $C(\Sigma)$, as the floor for $\alpha(\Sigma)$, and $|\Sigma|$ vary. The numerators possess the same property as elements of Pascal's Triangle.

$\alpha(\Sigma)$ > \ $\lvert\Sigma\rvert$ =	1	2	3	4	5	6	7	8	9
1	0	$\frac{1}{4}$	$\frac{4}{8}$	$\frac{11}{16}$	$\frac{26}{32}$	$\frac{57}{64}$	$\frac{120}{128}$	$\frac{247}{256}$	$\frac{502}{512}$
2	0	0	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{16}{32}$	$\frac{42}{64}$	$\frac{99}{128}$	$\frac{219}{256}$	$\frac{466}{512}$
3	0	0	0	$\frac{1}{16}$	$\frac{6}{32}$	$\frac{22}{64}$	$\frac{64}{128}$	$\frac{163}{256}$	$\frac{382}{512}$
4	0	0	0	0	$\frac{1}{32}$	$\frac{7}{64}$	$\frac{29}{128}$	$\frac{93}{256}$	$\frac{256}{512}$
5	0	0	0	0	0	$\frac{1}{64}$	$\frac{8}{128}$	$\frac{37}{256}$	$\frac{130}{512}$
6	0	0	0	0	0	0	$\frac{1}{128}$	$\frac{9}{256}$	$\frac{46}{512}$
7	0	0	0	0	0	0	0	$\frac{1}{256}$	$\frac{10}{512}$
8	0	0	0	0	0	0	0	0	$\frac{1}{512}$
9	0	0	0	0	0	0	0	0	0

Appendix D:

The respective sets, $\Delta_{n|p}$, which could be generated by a procedure of the following kind will be of dilution n, of pseudo level for dilution preservation p, and, as I have claimed, of level 2.

That is, $\forall n > 2, \forall p \geq n$: $\mathcal{d}(\Delta_{n|p}) = n$ & $\ell^{\mathcal{d}}(\Delta_{n|p}) = p$ & $\ell(\Delta_{n|p}) = 2$.

Schema 7: The schema functions in the same manner as schema 5, by listing the names of the sentences which will compose $\Delta_{n|p}$, and then describing the sentences.

For all $\Delta_{n|p}$ we stipulate the size of $\Delta_{n|p}$ to be p.

The elements of $\Delta_{n|p}$ are named $\alpha_1,..., \alpha_p$.

The sentences composing $\Delta_{n|p}$ are described as conjunctions generated with reference to a schema for generating n-membered sets of dilution n. We can use schema 4.

We take the first $\binom{p-1}{n-1}$ elements of an ordered set of *variations* of respective instances of schema 4. All variations of each instance are assumed to be pairwise disjoint with respect to their atomic subwffs. The elements of each variation are also assumed to be ordered.

Next we generate a set of sets $\Delta^{*n|p} = < a_{1_{n|p}}, ..., a_{P_{n|p}} >$ according to an algorithm which assigns a respective variation of schema 4 to each element of an ordering of the set of n membered subsets of $\Delta^{*n|p}$ which include a_1. Moreover, the *i*th element of the respective ordered variation of schema is assigned to the *i*th element of the ordered n-tuple.

Finally the set $\Delta_{n|p} = < \alpha_1, ..., \alpha_p >$ is generated by a procedure which assigns to each member of $\Delta_{n|p}$, a conjunction which corresponds to the elements of the corresponding elements of $\Delta^{*n|p}$.

Let $\Delta_{n|p}$ be defined as above then:

Theorem 29: $\forall n > 2, \forall p \geq n$: $\mathscr{d}(\Delta_{n|p}) = n \ \& \ \ell^{\mathscr{d}}(\Delta_{n|p}) = p \ \& \ \ell(\Delta_{n|p}) = 2$.

Proof of Theorem 29 (in three parts):

[1] $\forall n > 2, \forall p \geq n$: $\mathscr{d}(\Delta_{n|p}) = n$. The proof is straightforward. The instances of schema 4 which are used to construct $\Delta_{n|p}$ are disjoint with respect to their atoms. Thus, $\forall a \subseteq \Delta_{n|p}$: $|a| < n \Rightarrow a \nvdash_{PL} \perp$. Moreover, $\forall a \subseteq \Delta_{n|p}$: $|a| = n \Rightarrow a \vdash_{PL} \perp$.

[2] $\forall n > 2, \forall p \geq n$: $\ell^{\mathscr{d}}(\Delta_{n|p}) = p$. We need only notice that every pair of elements of $\Delta_{n|p}$ are members of an n-membered inconsistent subset of $\Delta_{n|p}$. Thus any aggregation will decrease the dilution of the set. Thus the pseudo-level of the set is equal to its size, which is p.

[3] $\forall n > 2, \forall p \geq n$: $\ell(\Delta_{n|p}) = 2$. All subsets of $\Delta_{n|p}$ which do not include α_1 are consistent. Thus for all instances of $\Delta_{n|p}$ there is a 2-partition of the set where both subsets are consistent. One cell contains α_1 and the other contains $\Delta_{n|p} - \{\alpha_1\}$. This concludes the proof.

Appendix E:

Proof of Lemma 3:

The lemma follows from the fact that $\binom{p}{n-1} - p + n - 1 = \left(\binom{p}{n-1} - \left(\frac{(n-1)\binom{p}{n-1}}{\binom{p}{n-2}} \right) \right) + 1.$

And from the fact that $\left(\binom{p}{n-1} - \left(\frac{(n-1)\binom{p}{n-1}}{\binom{p}{n-2}} \right) \right) + 1 =$

$$\min\{ g \mid \forall i \in \{2,\ldots,(n-1)\}: \left(g - (i-2)\left(\binom{p}{n-1} - g \right) + \left(\frac{\binom{n-1}{n-i}\binom{p}{n-1}}{\binom{p}{n-i}} \right) \right) > \binom{p}{n-1} \}.$$

Finally, it can be observed that

$\min\{ g \mid \forall i \in \{2,\ldots,(n-1)\}: \left(g - (i-2)\left(\binom{p}{n-1} - g \right) + \left(\frac{\binom{n-1}{n-i}\binom{p}{n-1}}{\binom{p}{n-i}} \right) \right) > \binom{p}{n-1} \}$ is a contorted version of what is explicitly demanded by the definition of $g(\Sigma)$.

Bibliography:

Apostoli, Peter and Brown, Bryson. *A Solution to the Completeness Problem for Weakly Aggregative Modal Logic.* Journal of Symbolic Logic. Vol. 60 (1995). pp. 832 -842.

Belnap, N. *How a Computer Should Think.* in Contemporary Aspects of Philosophy. (Gilbert Ryle, editor), Oriel Press, London (1976). pp. 30 -56.

Blackburn, Simon. *Theory, Observation, and Drama.* in Folk Psychology. (ed. Davies T., M. & Stone), Blackwell Pub. Ltd., 1995.

Botterill, G. *Folk Psychology and Theoretical Status.* in Theories of Theories of Mind. (ed. Carruthers, P. & Smith, P.K.), Cambridge University Press, 1996.

Brown, Bryson. *The Force of 2/n+1.* in Vicinae Deviae: Essays in Honor of R.E. Jennings. Simon Fraser University, 1993.

Carruthers, P. *Simulation and Self: a defence of theory-theory.* in Theories of Theories of Mind. (ed. Carruthers, P. & Smith, P.K.) Cambridge University Press, 1996.

Carruthers, P. & Smith, P.K. Theories of Theories of Mind. Cambridge University Press, 1996.

Cherniak, Christopher. Minimal Rationality. MIT Press, 1986.

Child, William. Causality, Interpretation, and the Mind. Clarendon Press, Oxford, 1994.

Churchland, P.M. *Folk Psychology and the Explanation of Human Behaviour.* in A Neurocomputational Perspective. MIT Press, 1992.

Churchland, P.M. A Neurocomputational Perspective. MIT Press, 1992.

Clark, Andy. Associative Engines. MIT Press,1993.

Collin, F. Theory and Understanding. Basil Blackwell, 1985.

Cummins, Robert. The Nature of Psychological Explanation. MIT Press, 1983.

Davidson, Donald. *Mental Events.* (1970) reprinted in Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, Donald. *The Material Mind.* (1973a) reprinted in Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, Donald. *Radical Interpretation.* (1973b) reprinted in Inquiries into Truth and Interpretation. Oxford: Clarendon Press, 1984.

Davidson, Donald. *Psychology as Philosophy.* (1974a) reprinted in Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, Donald. *Belief and the Basis of Meaning.* (1974b) reprinted in Inquiries into Truth and Interpretation. Oxford: Clarendon Press, 1984.

Davidson, Donald. *Thought to and Talk.* (1975) reprinted in Inquiries into Truth and Interpretation. Oxford: Clarendon Press, 1984.

Davidson, Donald. *Paradoxes of Irrationality.* in Philosophical Essays on Freud. Cambridge University Press, 1982.

Davidson, Donald. *Division and Deception.* in Actions and Events. (edited by E. LePore & B. McLaughlin.) Oxford, Blackwell, 1985a.

Davidson, Donald. *Incoherence and Irrationality.* in *Dialectica*, vol. 39, 1985b.

Davidson, Donald. *A Coherence Theory of Truth and Knowledge.* in Truth and Interpretation. (edited by Ernest LePore) Oxford: Blackwell, 1986.

Davidson, Donald. *Three Varieties of Knowledge.* in A.J. Ayer. Memorial Essays. (edited by P. Griffiths) Royal Institute of Philosophy, vol. 30, 1991.

Davidson, Donald. *Can There be a Science of Rationality?* International Journal of Phil. Studies, 1995.

Davies, M. & Stone, T. Folk Psychology. Blackwell Pub. Ltd., 1995.

Davies, M. & Stone, T. Mental Simulation. Blackwell Pub. Ltd., 1995.

Dennett, Daniel. *A Cure for the Common Code.* in Brainstorms. Bradford books, 1978.

Dennett, Daniel. *Reply to Professor Stich.* in *Philosophical Books,* 21, 2 (1980).

Dennett, Daniel. *True Believers.* (1981) Reprinted in The Intentional Stance. MIT Press, 1987.

Dennett, Daniel. *Three Kinds of Intentional Psychology.* (1981) Reprinted in

The Intentional Stance. MIT Press, 1987.

Dennett, Daniel. *Making Sense of Ourselves.* (1982) Reprinted in The Intentional Stance. MIT Press, 1987.

Dennett, Daniel. *Styles of Mental Representation.* (1983) Reprinted in The Intentional Stance. MIT Press, 1987.

Dennett, Daniel. The Intentional Stance. MIT Press, 1987.

Dennett, Daniel. *Real Patterns.* The Journal of Philosophy, 1991.

Dretske, Fred. Explaining Behaviour. MIT Press, 1988.

Evnine, Simon. Donald Davidson. Stanford University Press, 1991/

Fodor, J. The Language of Thought to. Harvard Univ. Press, 1975.

Fodor, J. *Fodor's Guide to Mental Representation.* in A Theory of Content and Other Essays. MIT Press, 1990.

Fodor, J. The Elm and the Expert: Mentalese and its Semantics. MIT Press, 1994.

Goldman, Alvin. *Epistemic Folkways and Scientific Epistemology.* in Liaisons. MIT Press, 1992.

Goldman, Alvin. *Interpretation Psychologized.* in Liaisons. MIT Press, 1992.

Gopnik, A. *How we know our minds: The illusion of first-person knowledge of intentionality.* (And open commentary) in Behavioural and Brain Sciences. Volume 16:1, March, 1993.

Gopnik, A. & Astington, J.W. *Children's Understanding of Representational Change...* in *Child Development* 59, 1988.

Gopnik, A. & Wellman, H.M. *Why the Child's Theory of Mind Really is a Theory.* (1992) Reprinted in Folk Psychology. (ed. Davies T., M. & Stone), Blackwell Pub. Ltd., 1995.

Gordon, R. *Folk Psychology as Simulation.* (1986) Reprinted in Folk Psychology. (ed. Davies T., M. & Stone), Blackwell Pub. Ltd., 1995.

Heal, Jane. *Replication and Functionalism.* (1986) Reprinted in Folk Psychology. (ed. Davies T., M. & Stone), Blackwell Pub. Ltd., 1995.

Heil, John. *Being Indiscrete.* in The Future of Folk Psychology. (edited by J. Greenwood.) Cambridge U. Press, 1991.

Jennings, R.E. Leibnizian Semantics. (Unpublished), 1984.

Jennings, R.E. and Schotch, P.K. *The Preservation of Coherence.* Studia Logica. Vol. 43 (1984). pp. 89-106.

Johnson-Laird, P.N. Human and Machine Thinking Lawrence Erlbaum Assoicates, Publishers, Hillsdale, New Jersey, 1993.

Kuhn, Thomas. The Structure of Scientific Revolutions. University of Chicago Press, 1962.

Kyburg, Henry. *Conjunctivitis.* in Induction, Acceptance, and Rational Belief. (edited by Marshall Swain.) D.Reidel Publishing Company, Dordecht-Holland, 1970.

Lakatos, Imre. The Methodology of Scientific Research Programs. Cambridge University Press, 1978.

Locke, John. Essay Concerning Human Understanding. published in part in Personal Identity. (edited by John Perry.) U. of California Press, 1975.

Malcolm, Norman. Dreaming. Routledge, London, 1962.

Moses, L.J. & Flavell, J.H. *Inferring False Beliefs from Actions and Reactions. Child Development* 61, 1990.

Morse, S.J. *The "Guilty Mind:" Mens Rea.* in The Handbook of Psychology and Law. (editors D.K. Kagehiro and W.S. Laufer) Springer-Verlag, 1992.

Perner, J. et al. *Three Year Olds' Difficulty Understanding False Belief. British Journal of Developmental Psych.* 5, 1987.

Priest, G. *The Logic of Paradox.* Journal of Philosophical Logic. Vol. 8. pp. 219 -241.

Ramberg, B. *Naturalizing Idealizations: Pragmatism and the Interpertivist Strategy.* (Draft, 1996)

Reichenbach, Hans. The Philosophy of Space and Time. Dover Publications Inc, New York, 1958.

Rescher, N. and Brandom, R. The Logic of Inconsistency. Blackwell, 1980.

Routley, R. and V. *The Semantics of First Degree Entailment.* Noûs. Vol. 6. (1972). pp. 335 - 359.

Ryle, Gilbert. The Concept of Mind. Barnes and Noble, New York, 1963.

Schotch, P.K. and Jennings, R.E. *Inference and Necessity.* Journal of Philosophical Logic. Vol. 9 (1980). pp. 327 - 340.

Schotch, P.K. and Jennings, R.E. *On Detonating.* in Paraconsistent Logic. (G. Priest and R. Routley, editors), Philosophia Verlag, Munich (1989), pp. 306-327.

Sellars, Wilfred. *Empiricism and the Philosophy of Mind.* in Science, Perception and Reality. Ridgeview Publishing Company, 1963.

Stich, Stephen. *Headaches.* in *Philosophical Books,* 21, 2 (1980).

Stich, Stephen. *Dennett on Intentional Systems.* Philosophical Topics, 12(1), 1981.

Stich, Stephen. From Folk Psychology to Cognitive Science. MIT Press, 1983.

Stich, Stephen. Deconstructing the Mind. Oxford University Press, 1996.

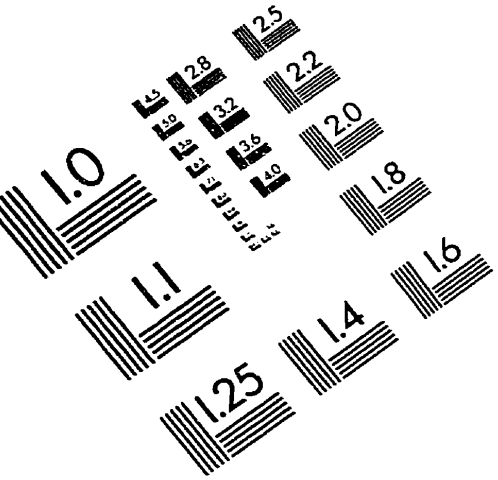
Wellman, H.M. The Child's Theory of Mind. MIT Press, 1990.

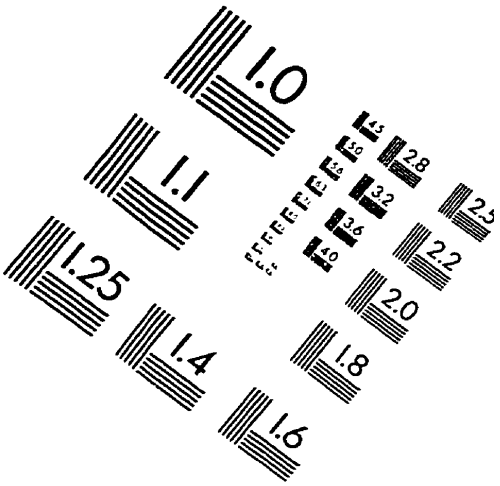
Wimmer & Hartl. *The Cartesian View and the Theory View of Mind. British Journal of Developmental Psych.* 9, 1991.

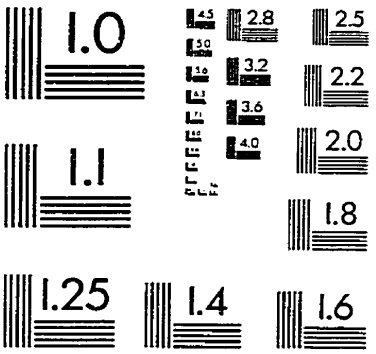
Wittgenstein, Ludwig. Philosophical Investigations. (Translated by G.E.M. Anscombe), Prentice Hall, 1958.

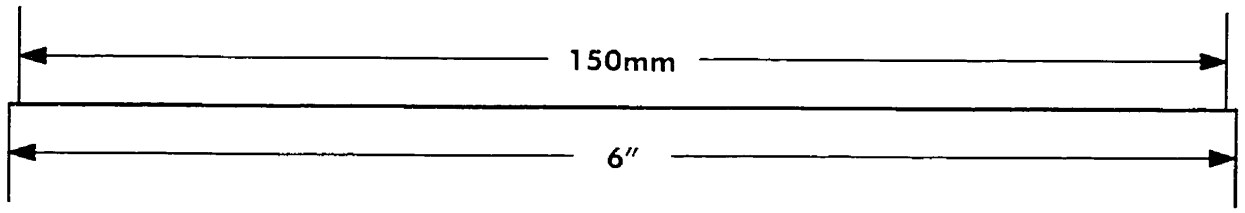
Wittgenstein, Ludwig. The Blue and Brown Books. Oxford: Blackwell, 1964.

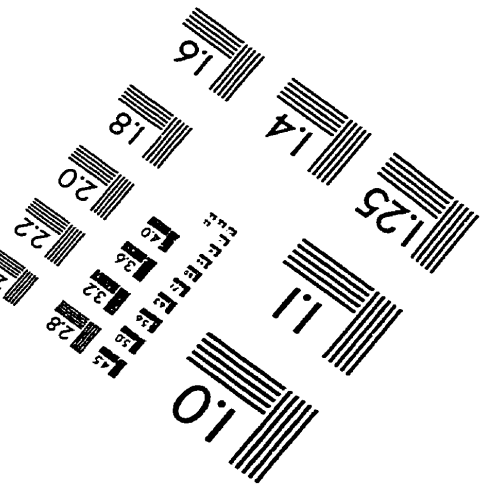
IMAGE EVALUATION TEST TARGET (QA-3)

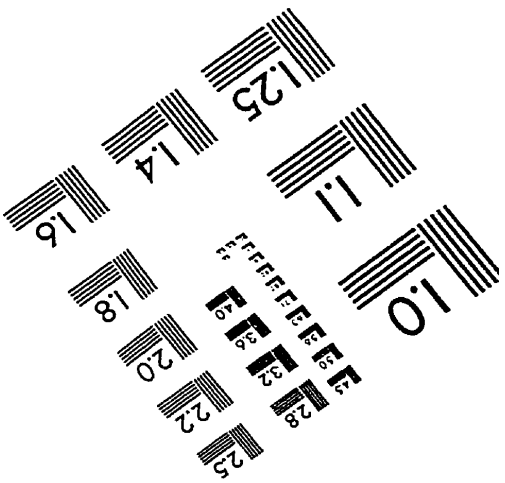












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989