In Praise of Wider Functionalism
or
For more matter in mind

By J. S. Edwards

Ph. D.
University of Edinburgh
1987
Declaration:

This thesis has been composed by myself and is entirely my own work.

signed

J. S. Edwards
The premiss of this thesis is naturalism: viz., that psychological facts whilst not reducible to are determined by physical facts. Pace Davidson's anomalism of the mental, the Author argues that token-token psycho-physical identity presupposes that inferential inter-relations between psychological states are projectably mirrored by physical dispositional relations between those states.

The Author aims to regiment our folk-psychological explanations because he argues, we have warrant to apply the conceptual contents of our self-applied indices to the states so indexed, rather than treating those conceptual contents merely as indices of internal states. That warrant arises from the determinate conceptual content of our home language - pace the well-known arguments of Quine, Putnam and Davidson.

The Author also aims to show that only a functional system wide enough to include the objects which we perceive and manipulate can generate those conceptual contents, not a narrow functional system which terminates at or before the bodily envelope. Field's subjective semantics for classical logic are used to show that narrow functionalism cannot account for the sense of "I", nor, adapting work of Hütikka and Rantala, for our quantifier rules, and narrow functionalism leads to brute psycho-physical correlations between sensational properties and inner physical states. Fodor's formality condition does not preclude mental operations having access to the semantic contents of mental states. And the Author argues for a wider functionalism which is blind to the particular identities and differences of res, but is not blind to their observable properties.

The Author develops an account of quantifying into attitudes and quasi-binding variables in content clauses without inducing transparency, thus enabling individual attitudes to be linked up and mutually oriented in action explaining complexes. Such quantification is extensionally defined so as to fit an otherwise standard truth definition. (Kripke's puzzle about Pierre's beliefs is thereby resolved.)

The Author shows how any individual episode in a person's psychological history can be treated narrowly, as an hallucination, without explanatory loss, but provided enough of his episodes are explained widely, quantifying over res, sufficient to account for the conceptual content of all episodes.
CONTENTS

Chapter One: A sketch of the relation between the mental and the physical ................................ 1
(i) The mental and the physical ................................ 1
(ii) Functionalism alone does not imply the irreducibility of the mental to the physical ... 6
(iii) Psychological laws, psycho-physical laws and mental-physical identity ....................... 9
(iv) Points of agreement and disagreement with Anomalous Monism summarised .......... 23
Appendix A: Field's semantics for catastrophically vague terms ....................................... 25
Appendix B: The supervenience of the psychological upon the physical ......................... 30
Footnotes to chapter one ........................................................................................................ 34

Chapter Two: Conceptual content and psychological realism ............................................... 36
(i) Introduction ........................................................................................................................ 36
(ii) A criterion for psychological realism proposed ......................................................... 41
(iii) Does narrow functionalism fail the test for psychological realism? ......................... 45
(iv) The psychological status of dumb animals ............................................................... 49
(v) The psychological status of fellow language users .................................................... 61
(iv) An assertion-belief generalisation ............................................................................... 70
Appendix A: Quine, Davidson and psychological realism .................................................. 75
Appendix B: Kripke's doubts about his disquotation schema ............................................ 82
Appendix C: McGinn on the relation between reference and belief .................................. 96
Footnotes to chapter two ........................................................................................................ 101
Chapter Three: On the alleged indeterminacy of conceptual content .......................... 105
   (i) Introduction .................................................. 105
   (ii) Putnam on the indeterminacy of reference .................................. 108
   (iii) Davidson's sub-sentential instrumentalism .................................. 119
   (iv) Quine's Duhemian argument ............................................. 132

Appendix A: Empirical equivalence and ontological equivalence ......................... 150

Footnotes to chapter three .................................................. 155

Chapter Four: Narrow and wider functionalism ............................................. 158
   (i) Introduction .................................................. 158
   (ii) Fodor's narrow functionalism .......................................... 159
   (iii) Narrow functionalism and the 'machine code' .......................... 166
   (iv) Narrow functionalism and rational inference .......................... 171

Appendix A ................................................................. 197
Appendix B ................................................................. 203

Footnotes to chapter four .................................................. 209

Chapter Five: De re attitudes and psychological explanations .......................... 211
   (i) Introduction .................................................. 211
   (ii) The role of de re attitude reports in psychological explanations. And the role of de re attitudes in the causation of action ............................. 214
   (iii) Utterance meaning and the diversity of individual functional psychological theories .......................................................... 226
   (iv) Quantification-in and the truth definition .................................. 237
   (v) Hallucination ..................................................... 243
   (vi) Peacocke on modes of presentation ...................................... 252

Footnotes to chapter five .................................................. 259
Summary of the main contents of each chapter

Chapter One

The Author argues from naturalism (= psychological facts are determined by but not reducible to physical facts) and monism to the conclusion that cognitive and volitional states are second-order physical dispositional states - where the de facto dispositional inter-relations projectibly mirror the de jure inferential inter-relations of the psychological states, so as to determine psycho-physical identities. The argument proceeds by criticism of Davidson's Anomalous Monism: there the identity is with first-order physical states, and the resulting anomalism, it is argued, vitiates the attendant claim of monism.

Chapter Two

Indexing Tom's internal states by sentences leaves open the question whether or not those internal states are representations and have the conceptual contents of their indices. The Author proposes that they are if and only if, where there are two systems of sentential indices assigning different conceptual contents to Tom's internal states, the corpus of physical facts about Tom can be enlarged so as to eliminate one system of sentential indices - since, if the indices are to show what it is like (cognitively and volitionally) to be Tom, they must show him to be of one mind.

The Author doubts whether narrow functionalism can pass this test: the belief indexed by, say, "Boise is a city" could, it seems, also be indexed by "'Boise' is stored at the address CITY", a sentence with a different conceptual content. The Author also doubts whether animal behaviour passes this test - because assigning conceptual content presupposes grouping behaviour by type; and non-verbal behaviour cannot, whereas verbal behaviour can (more or less) be grouped by us into relevant type prior to our assigning to those types
intentional (conceptual) contents.

Where attitude reports are in Tom's home language, then an assertion-belief generalisation of the rough form

Tom believes that $p$ iff Tom is disposed to assert that $p$ has an a priori status, provided the sentences of Tom's home language have determinate conceptual contents. We can take the attitudes which Tom is thus disposed to avow as forming one psychological history, and thus Tom automatically passes the test for applying the conceptual contents of those self indices to his internal states so indexed.

In Appendix B Kripke's worries about his related disquotation schema lead to a distinction between syntactically inconsistent beliefs (which functionalism cannot handle) and semantically inconsistent beliefs (which it can handle). And this distinction leads on to the idea of quantifying into attitude reports so as to quasi-bind variables in the content clauses, whereby semantically inconsistent beliefs can be handled by syntactically defined inference rules in such a way that the behaviour so indexed is shown to be rational.

Chapter Three

The Author is committed to the reference of open and closed sentences being scrutable and their conceptual contents being determinate.

Putnam's argument to the contrary is stripped of its unhelpful detour through metalanguages to highlight the central claim that physical facts (narrow or wide) fix reference (if they do) only via fixing truth conditions. However, an examination of Davidson's work shows, pace Davidson, that a theory of meaning for a language containing indexicals needs needs to employ a semantic conventional relation of reference in the observation sentences of the theory, whence such a relation must be deemed real and scrutable. So, although there is no route from truth conditions to a unique assignment of
referents, physical facts determine truth conditions only via determining conventional referential relations.

Quine’s ‘Duhemian’ argument fails because ‘observation’ terms penetrate to the heart of a theory. In this way we balance two opposing pulls of naturalism: that meaning should be manifest in observable use, the senses of theoretical sentences being controlled by that of observation sentences, and that theoretical entities are not ontologically dependent upon those entities which happen to be observable to us. (Quine’s attempt to show that empirically equivalent theories are really ontologically equivalent is rejected because it leads to an unworkable brand of holism.)

Chapter Four

Narrow and wider functionalisms are distinguished. Fodor’s narrow functionalism supposes that inference rules preserve designated semantic values however the creature is embedded in whatever possible world.

Fodor’s formality condition does not imply narrow functionalism - provided one takes a relational view of the syntax of MENTAL. Fodor implicitly recognises this in the case of a machine code, where the relevant functional system includes the semantic contents. The sentences of machine code (narrow functionalism) need not be the sentences of a wider functional system which they manipulate.

Field’s semantics for classical logic using conditional subjective probability provides a model for narrow functionalism. But it fails to give “∃” the sense of “there is” rather than “there may be”, with variable modality. To disambiguate we need to see how the subject’s beliefs of the form (3x)Fx relate to its environment - to see whether reading “∃” as “there is” makes him rash of judgement, or reading “∃” as “there may be” makes him overcautious.

Hintikka and Rantala’s work shows that whether classical
quantifier rules preserve designated semantic values
depends upon the wider functional system, how the subject
is embedded in what possible world. Inductive strategies
also presuppose a user-friendly world.

A narrow functionalist cannot drop his requirement
that computations preserve designated semantic values
without losing his claim to be indexing the relevant
internal functional states.

In Appendix B the Author recommends that functionalists
do not try to incorporate syntactically inconsistent
beliefs. Rather, functionalists should classify syntactic
inconsistency as malfunction, as a discontinuity in the
subject's psychological life. (Loar's minimal standard
of rationality is too weak to sustain a computational
chain.)

Chapter Five

The Author argues that quantifying into attitude
reports to quasi-bind variables links up and mutually
orients the individual attitude reports into an explanatory
complex - a complex upon whose content clauses syntactically
defined inference rules operate so as to demonstrate
the rationality of the agent. And the attitudes so reported
are de re: the agent's judgement that the object of one of
his attitudes is the same as or is different from the
object of another is as primitive as, to consciousness,
the judgement that they share a common observable property.
The conscious warrant for such 'bare' judgements of
identity or difference is holistic: the success of the
resulting belief system. Quasi-binding is sufficient to
link up and mutually orient content clauses without
inducing transparency. The Author gives an extensional
definition of quantifying and quasi-binding variables
which may be fitted into an otherwise standard definition
of truth.

Taking a theory of meaning to assign truth conditions
and a syntactic structure to each sentence, and taking
sentences with the same truth conditions to differ in
syntactic structure iff it is possible to believe that the one and not that the other, then it follows that the rules \( \vdash \) operating on the content clauses of attitude-reports are rules for manipulating syntactic structure such that \( S_p = S_q \iff p \vdash q \). And the resulting functional theory is really the description of a family of theories realised by the various members of the one language community.

If quantification-in is needed to form an explanatory complex, then Tom, a dagger-hallucinator, cannot be in the same psychological state as Twin Tom, a dagger perceiver. Yet it seems to Tom as it seems to his twin because they share the same psychologically relevant physical inner states. An explanatory complex is sketched for hallucinating Tom.

Peacocke's account of modes of presentation does enable the agent's current attitudes to be linked into an explanatory complex. But diachronic rationality, the linking of attitudes across time, still requires 'bare' judgements of identity and difference of res.

Chapter Six

The wide functional system is blind to the identities and differences of the res of attitudes, but is it also blind to the observable properties of those res? A narrow functionalist must suppose so, but a wider functionalism can, it seems, go either way.

The Author argues that Putnam is correct to suppose that Tom and Twin Tom share a common concept WATER, but not because that concept contains a concealed indexical. (For usage does not manifest an indexical waiting to deal with a mere fantasy.) Rather, the different extensions, water and twater, are determined by different overt communicative intentions.

If the wide functional system quantifies over observation properties, then SHRDLU (the block-manipulating robot) and T-SHRDLU (the computer manipulator) share the same observation concepts, and whether an object instantiates
a given concept or not is not mind-independent but depends upon a cognitive viewpoint - as does whether WATER applies to water or to twater.

However, this approach fails for observation concepts. For we would be forced to define the common observation concepts in terms of *sensational* properties, and, since SHRDLU and I-SHRDLU (the robot with the spectrum-inverting lens) realise the *same* wide physical functional system and therefore the *same* psychological history, it becomes arbitrary which sensational property is correlated with which physical state, which offends against naturalism. Hence the Author requires the psychologically relevant physical facts to include the observable properties, which can then be employed to define, non-arbitrarily, the corresponding sensational properties. Narrow functionalists cannot avoid arbitrary psycho-physical correlations. (Fodor's reasons for pessimism about the prospects for a wide functionalism do not apply to observable properties.)

**Appendix A**: Lewis, a narrow functionalist, recognises the possibility of an inverted spectrum, but cannot say what it is like to be the victim of such.

**Appendix B**: The Author is committed to those observable properties which are included in the psychologically relevant physical functional theory being primary, not secondary qualities. The Author rejects McGinn's arguments that the primary-secondary quality distinction is a priori. The author argues that we are biased towards primary qualities, but that the circumstances which would force a decision have not yet arisen, and should they do so it is an a posteriori matter which way our language community would go.

**CONCLUSION**

Drawing the treads of the previous discussion together, the Author concludes that whilst any episode in the psychological life of Tom can be treated narrowly as an hallucination, without explanatory loss, not all can together.
Chapter One: A sketch of the relation between the mental and the physical

What is the relation between the mental and the physical; more particularly, between cognitive and volitional states and events on the one hand and on the other neurological states and events? In this chapter I shall outline an answer to this question - an answer at the same level of generality as "Each token mental state or event is a token physical state or event". I shall not, in this chapter, attempt to justify every feature of my answer, but only those features which differ from other well-known answers: viz. anomalous monism and functionalism. This is partly because I have nothing new to add, at this level of generality, in justification of those features it shares with anomalous monism and functionalism; also partly because the plausibility of such a general answer depends upon how well it coheres with one's metaphysical prejudices, in my case for naturalism and against Cartesian dualism, whereby the answer reinforces those prejudices and they, in turn, render the answer plausible; and partly because, I hope, my answer to this central question in philosophy will become more plausible in later chapters, when we descend from Olympian generality sufficiently to notice other issues - viz. the semantic content and sense of representations (wide or narrow functionalism), the role of de re attitudes in psychological explanations, and related issues.

Section (i): The mental and the physical

Here, then, is my general answer to the question: How are cognitive and volitional states and events related to neurological states and events? Firstly, the concepts employed in belief-desire psychology and the psychological facts they are used to report are not reducible to physiological concepts
and facts. This is because of the normative nature of psychological explanations, because beliefs and desires justify an action insofar as they show that action to be rational, whereas physical facts, even dispositions to bodily motion in response to stimuli, lack normativeness; and also because psychological explanations involve representational properties of beliefs and desires. Let us suppose, as has been forcibly argued, that psychological explanations involve only the senses of representations and not their semantic properties (narrow functionalism), though I shall reject this supposition in chapter four, nonetheless, there seems no prospect of defining or reducing senses to physical dispositions, and this again because sense involves the normative idea of evidence justifying or not justifying belief or assertion.

Secondly, I take belief-desire psychology to be a functional theory: that is to say, a belief is the particular belief it is because of its normative relations to other beliefs, desires, perceptions and intentions. (This makes the phenomenon of irrational belief philosophically problematic: a problem I explore in chapter two, Appendix B, and in chapter four, section (iv) and appendix B.) I take a subject’s neurological states and events, suitably inter-related, to realise the beliefs, desires, etc. which the psychological theory attributes to him.

Functional theories are either theories about the inter-relations of first-order entities, or theories about the inter-relations of second-order entities, where those second-order entities are themselves constituted by some
first-order entities. Since the distinction will prove crucial in what follows, let me illustrate it at a little length.

My theory of my motor car is a first-order functional theory: the parts of my car are concrete objects which are only very contingently causally related as the theory requires - my car may have a different offside rear wheel today from yesterday, even though they each fulfil the same function, and the old wheel is still a wheel even though it now only gathers rust in my garage, and if two entities perform the same function, e.g. dual braking systems, then I have two braking systems. On the other hand, my theory of the telephone is a second-order functional theory. Calls are charged by duration, distance and time of day. I may note that I made five calls during the morning. It may be that one and the same call is switched from one channel to another as traffic on the network ebbs and flows, and it may be that at some times my call is constituted by more than one channel simultaneously, each individually sufficient to carry the call, perhaps because it is electronically better to keep channels occupied then to let them lie idle. My subscriber's theory is a functional theory quantifying over calls, second-order entities relative to channels, and an engineer might know what channels contingently constituted which of my calls.

Now, beliefs, desires, perceptions and intentions are, I think, second-order entities relative to the neurological first-order states and events whose contingent causal interrelations constitute the psychological states. For, if beliefs etc. were first-order neurological states or events, then, since the causal relations of first-order states and
events are contingent, my belief that p might not have had that representational content and might have not been a psychological state at all. And if, as seems likely in view of the mortality of nerve cells, my belief state was realised first by one neural circuit and later by another, then it would not be the same token belief, but two beliefs of the same type. And similarly, if my settled belief that q is multiply realised in my neural circuitry so as to be failsafe barring catastrophic neural damage, then an action was not caused by my belief that q, as commonsense has it, but by my beliefs, all of the type that q.

I do not think that the mere strangeness of saying that my beliefs that p caused my action, or of saying that one of my beliefs that p might have been preserved in vitro as an isolated specimen in Dr. Frankenstein's laboratory, rules out a first order interpretation of the entities postulated in psychological explanations. Whether the theory is to be interpreted as quantifying over first or second-order states, may be simply undetermined by present usage, the question not having arisen in that usage.

However, I conjecture, if we did ever find one set of causally inter-related first-order states and events which realised a concrete psychological history, then there would be indenumerably many other ways of caving up that concrete causally-related whole into proper parts which would also realise the psychological history. And this militates against a first-order interpretation of psychological theory.

Perhaps, given a sufficiently pure and abstract theory
of the motor car, the theoretical entities could be multiply constituted by more than one inter-related system of first-order proper parts of one and the same concrete motor car. But we use the theory of the motor car when repairing and replacing parts, so that one particular first-order realization is salient to us, and is built in to the 'applied' theory; the beginner is taught that this is one carburettor, that is one petrol pump, etc. Thus we are motivated to learn a first-order theory of how a motor car functions. In contrast, we employ psychological theory in blissful ignorance of the relevant first-order neurological contents of skulls.

I conclude, then, that the first-order neurological events or states between which the relevant contingent causal relations hold are not identical with but rather constitute the second-order psychological states or events, where the relation of constitution is not identity but may be many-one. Psychological theory is a second-order functional theory.

I confessed preudice for naturalism and against Cartesian dualism. In the present context this amounts to a belief that, insofar as there are psychological facts of the matter, those facts, although not reducible to neurological or other physical facts, must somehow be determined by physical facts (naturalism), and beliefs, desires etc. must be identical with physical states (against Cartesian dualism), albeit second-order physical states. But for beliefs, desires, etc. to be second-order physical states there must be a second functional theory, a physical theory postulating dispositions whose functional inter-relations mirror those attributed to the beliefs, desires, etc. by the psychological theory.
Because the inter-relations between beliefs, desires, etc. attributed to a given subject by the psychological theory are mirrored by the inter-relations between the second-order physical states attributed to him by the physical functional theory; because of this, I claim, we can identify the token second-order states attributed by the one theory with the token second-order states attributed by the other, and claim that the physical functional facts realise the psychological facts (against Cartesian dualism and for naturalism), without maintaining that psychological concepts and the facts they report are reducible to the physical concepts and facts.

To say that second-order psychological facts are realised by second-order physical states is to say that each token psychological state is identical with such a token physical state. Of course, if the entities postulated by the psychological theory are second-order entities constituted by first-order entities, where the relation of constituting may be many-one, then the entities postulated by the physical dispositional theory, being the same entities, will be second-order and constituted, perhaps many-one, by first-order states.

That completes my answer to the question: What is the relation between mental (cognitive and volitional) states and physical states? Now to relate this answer to functionalist orthodoxy and to anomalous monism.

Section (ii): Functionalism alone does not imply the irreducibility of the mental to the physical.

My claim that psychological states are second-order functional states constituted by but not identical with
first-order physical states is no novelty: Putnam talks of machine states being "abstract" or "logical". But to avoid dualism, I have postulated that those same states are also the subject matter of a second, purely physical functional theory. Here I depart from Putnam: for Putnam locates the irreducibility of the psychological to the physical in the irreducibility of a functional theory to first-order causal laws. It is a recurrent theme in Putnam that the irreducibility of the mental to the physical is just one more case of the irreducibility of functional to first-order theories. No-one, I take it, would dispute that functional theories of motor cars or telephone systems are naturalistic, so Putnam is claiming that the alleged contrast between mental and physical theories and facts is misconceived. I differ from Putnam in that I see the contrast between the normative and the natural as real, rendering psychological concepts and facts irreducible to physical concepts and facts. However, the psychological history generated by applying normative psychological theory to Tom - the story of what perceptions produced what beliefs, and what beliefs and desires produced what actions - is mirrored by a physical, non-normative, dispositional theory applied to Tom: the state to state transformations reported by the one theory can be mapped one-to-one onto the state to state transformations reported by the other theory, where each state and its image state are constituted by the same first-order physical states, and this justifies our identifying the token states postulated by the one theory with the token states postulated by the other, or so I claim. Hence, if the warrant for claiming mental-to-physical irreducibility were
as Putnam claims, there would be no reason to claim irreducibility. A physical functional theory, pitched at the same level of abstractness and generality as the psychological theory, would be available.

I agree with Putnam that functional theories cannot be reduced to first-order theories, and this for the reasons Putnam gives - briefly, firstly that contingent first-order premisses, describing the contingent make-up of the machine, are needed in order to deduce the functional laws of its operation from the basic laws of physics, and secondly that the various concrete realisations of a functional theory do not form a natural kind: quite different first-order laws explain the operation of a cog-and-lever calculator from those explaining an electronic calculator, and there is no way of enumerating the possible realisations of a given functional theory in first-order terms, and so no way of deducing that theory from first-order premisses. My disagreement with Putnam is that I do not see the irreducibility of functional to first-first order theories as the genus of which the irreducibility of the psychological to the physical is a species.

Although a second-order physical functional theory applied to a particular concrete machine will provide a distinction between first-order changes which are functions from first-order changes which are malfunctions, these words should be divested of any normative content: the degeneration of my car into a heap of rust is just as natural a causal process as was its transporting me from A to B in happier days.
This may shed some light on the difficulty a Martian anthropologist may have in grasping the rules of human games, in knowing 'how to go humanly on'. Let him know all the first-order facts about humans, his problem may be that he doesn't know where to draw the line between function and malfunction, normal development and pathology, signal and noise. Tom's scratching his ear may be irrelevant so far as his play of the game is concerned, but perhaps not irrelevant to all psychological theory since it may manifest his irritation and frustration with the game. The interesting sense in which the Martian is an alien intelligence may be that no way seems salient to him of stacking the first-order facts about humans into second-order facts, so the Martian fails to spot the relevant physical second-order facts which realise the normative psychological game-play.

Section (iii): Psychological laws, psycho-physical laws and mental-physical identity

My answer is complexly related to anomalous monism: preserving, I hope, its insights whilst avoiding, I shall argue, its incoherence. Davidson finds three features characteristic of psychological explanation: firstly, psychological explanations are holistic, any concrete attribution of beliefs, desires, etc. is defeasible in the light of a wider context of beliefs, desires etc.; secondly, psychological explanations are normative, governed by the constitutive master concept of rationality - I strongly suspect that Davidson regards these two features as not mutually independent. And thirdly, and perhaps also
consequently, psychological generalizations are inherently inexact, hedged by *ceteris paribus* clauses, and this because of the nature of the subject matter and not because our psychological theorising is at some pre-Newtonian stage of under-development.

I have accepted the second feature, the normativeness of the psychological, and any form of functionalism incorporates the first feature, holism: a concrete machine cannot realise one internal state specified by a Turing machine-table unless it potentially realises the whole machine-table, which is to say that, given suitable sequences of inputs it would pass from that state to the others delivering up output as specified by the table. So the claim that the machine realises a given machine-table is defeasible, and would be defeated by further evidence that it failed to progress and respond as per the machine-table. (Of course, a concrete machine may be in state A at time $t_0$, and fail to be in B at $t_1$, though the table required it, because the machine has broken down between $t_0$ and $t_1$. But that is quite a different point: by $t_1$ it has ceased to realise the functional theory, and the defeasibility we are interested in is the defeasibilty of claims about the current state of a well-functioning machine; claims being made on the basis of its history of observable input and output.)

In certain cases, where the input tape is finite, and the machine-table is finite, it will be possible in principle to arrive at an *indefeasible* claim that the machine was in state A, by checking out all the related possibilities, a comprehensive run through all the machines functional
capacities. (Or rather, if such a claim remains defeasible that has nothing to do with it being a claim about a functional state, it being merely the defeasibility of any empirical claim about the world.) But such a possibility does not refute the holism of functional claims, it merely reflects that the whole in question was a finite whole. And, of course, in general, because the input tape is unlimited, the relevant whole will be infinite, and claims about the internal state of such a machine based on a finite history of observable input and output, will be defeasible in principle.

The third feature which Davidson sees as contributing to the anomalousness of the mental, the inherent inexactness of psychological generalisations, is also, when correctly interpreted, incorporated in functionalism. But, when correctly interpreted, it renders the resulting anomalism incompatible with Davidson's form of monism!

A functional theory may be as precise as we please: there need be no vagueness, no ceteris paribus clauses, in a Turing machine-table. But a concrete machine may realise the functional theory only intermittently: computers 'go down', cars have to be towed to garages, etc. A concrete machine realises the relevant functional theory only if certain 'boundary conditions' are maintained: serving and repair sustain the boundary conditions. The crucial point, for what follows, is that the boundary conditions cannot be expressed within the relevant second-order functional theory, except tautologically as the condition that the machine must be such as to realise the functional theory. The boundary
conditions cannot be expressed non-tautologically within a second-order functional theory because they vary from concrete first-order constitution to concrete first-order constitution, of the same functional theory. A mechanical calculator and an electronic calculator may both realise the same second-order functional theory, but their servicing manuals will be very different. And further, we recall, there is no way of listing the possible concrete realisations of a given second-order functional theory, and so no way of incorporating the different possible boundary conditions into the functional theory.

We can now see a sense in which psychological and other second-order functional laws are 'inexact' in contrast to an 'exactness' of physical first-order causal laws. It is not that psychological laws are only rough generalisations, imprecise or festooned with ceteris paribus clauses. They may be as precise and exceptionless as we please. But concrete machines only intermittently instantiate those laws: when they do not instantiate those laws, when the car breaks down, it does not falsify the theory of the motor car; rather, it ceases to be, for a while, subject matter for that theory. Now first-order causal laws may themselves be subject to boundary conditions: kittens fed on Katomeat develop into mature cats, provided they are not carried off by accident or disease. Davidson's point is, I take it, that in the case of the rough causal law "Kittens fed Katomeat grow into mature cats" we can refine and develop the law by adding specific first-order provisions until any ceteris paribus condition is discharged and we have an exact and
exceptionless causal law. That is to say, the relevant boundary conditions can be incorporated into the law so that the resulting law is of the form

$$(x)(Ax \rightarrow Bx)$$

where the domain of the quantifier is unrestricted. In contrast to this, although a second-order functional law of the above form may be as precise as we please, the domain of its quantifier is restricted to realisations of the functional theory, because the various boundary conditions cannot be incorporated into the functional law. So the domain is whatever entities satisfy the whole theory of which that law is one part.\(^7\)

These three features of belief-desire psychology are supposed by Davidson to render the mental anomalous. Davidson gives two significantly different explanations of his term of art: In one explanation\(^8\) the term refers to the lack of laws relating the mental to the physical, there may be generalisations, even perhaps exceptionless generalisations, but such would be merely heteronomic co-incidences of mental and physical properties, lacking the projectibility of laws. The second explanation formulates a "Principle of the Anomalism of the Mental: there are no strict laws at all on the basis of which we can predict and explain mental phenomena".\(^9\)

Davidson clearly holds that the Principle of the Anomalism of the Mental, together with the plausible premiss that all physical events are subject to strict laws, entails anomalousness understood now as the lack of psycho-physical laws. Here is a typical passage to that effect.\(^10\)
There are no strict psychophysical laws because of the disparate commitments of the mental and physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual. There cannot be tight connections between the realms if each is to retain allegiance to its proper source of evidence.

I have accepted the Principle of the Anomalism of the Mental: there are no strict laws on the basis of which we can predict and explain action. By a strict law I mean one whose boundary conditions, if any, can be incorporated into the law, non-tautologically, so as to render the universal quantifier unrestricted. So understood, psychological laws are not strict laws. Further, their use to explain some concrete past mental event, or to predict some future concrete mental event, will indeed be accompanied by the ceteris paribus proviso that the relevant boundary conditions were or will be fulfilled. And discharging that ceteris paribus proviso is no part of psychological theory, but is part of the first-order theory of the concrete creature in question. So understood, the Principle of the Anomalism of the Mental is merely an instance of the Principle of the Anomalism of second-order functional theories.

However, so understood, the principle does not entail anomalism understood now as a lack of mental-physical laws. The failure of functional psychological theory alone to completely explain a mental history or predict mental events is consistent with there being projectible psycho-physical laws mapping mental events onto second-order physical events,
physical events whose inter-relations according to the physical dispositional theory mirror the inter-relations of those mental events as recorded by the psychological theory.

Let us concede to Davidson that the status of such mental-physical correlations is mightily obscure at present, given that psychological facts and concepts are not reducible to dispositional facts and concepts because of the normativeness of the former and the matter-of-factness of the latter, and also given that those correlations need to be strong enough to support a claim that the mental events cited by the psychological theory are identical with the physical events cited by the dispositional theory, so as to avoid Cartesian dualism. (I hope to dispel some of the obscurity in later chapters.)

To try to convince my reader that the dispositional functional theory is needed, that monism is to be achieved by identifying psychological states with second-order physical states guided by some projectible relation between psychological and dispositional states, I propose to show the difficulty which anomalism generates for Davidson's monism, a monism in which psychological states are identified with first-order physical states, and in which the identifications are not guided by any projectible psycho-physical relations.

Davidson's monism is based on the premiss that any causal relation between states or events must be covered by a strict law, where by a strict law we mean, as explained above, one in which any boundary conditions can, in principle, be so incorporated into a revised version of the law that the domain of the resulting quantifiers is unrestricted.
And psychological laws, because they are second-order functional laws, do not meet this condition. Davidson's other premiss is that perceptions, beliefs, desires and intentions partake in causal relations. (I reject this premiss, holding instead that the events between which the causal relations hold are not identical with but constitute the mental events, where constitution can be a many-one relation. Mental events are active in the physical world by virtue of being constituted by physical events which are themselves causally interactive.) From these premisses Davidson concludes that mental events are identical with those causally interactive physical events.

If mental events or states are physical events or states, then the following is true, where "b" names my belief that the moon is made of green cheese and "Px" means "x is a physical event or state":

\( \exists x \) (Px & x=b)

An anti-realist will require, for the above to be true, that we can find, in principle, some value of the variable such that we are warranted in asserting something of the form

Fa & a=b

since anti-realists require that, in all cases, "(\exists x) (Fx)" be assertible only if some sentence of the form "Fa" is assertible. And further, a realist will hold that for the quantified sentence to be true, there must be an answer to the question "Which physical event or state is identical with that belief?", whether we can determine that answer or not, Identity is a one-one relation, so if a is not the same
physical event as c, then, for a realist, "a=b" and "c=b" must have opposite truth values, and for an anti-realist they cannot have the same truth value. On neither view is it up to us to stipulate or choose what their truth values shall be. Let us call this the Principle of the Determinateness of Identity (PDI).

Having stated PDI we must blunt its cutting edge somewhat by recognising various sorts of vagueness in our sortal concepts. Arguably, this lump of rock is identical with a collection of molecules, rather than being constituted by that collection of molecules. But it is indeterminate which collection of molecules, precisely, is this lump of rock: molecules are lost and gained by the collection all the time and our usage does not manifest precise boundaries because we have had no need to determine precise boundaries. However, this unmysterious vagueness, in our conception of a lump of rock relative to our conception of a collection of molecules, does not undermine the claim of identity, just because any two candidate collections massively overlap each other (although no molecule need be common to all candidate collections).

Our concept of action manifests another species of tolerable vagueness. Some philosophers have claimed that actions are physical events, but have differed amongst themselves over which physical events they identify with actions: one offers purely internal neural events, another these plus the bodily motions they cause, another the bodily motions, and yet another the bodily motion plus the resulting causal chain out to and including the intentional object
I find the resulting rival polemics unconvincing. Any choice of candidate leaves hard cases: for example, if Claudius' killing of Hamlet ceased with Claudius' bodily motion then he had killed Hamlet before Hamlet was dead because, like other Shakespearean heroes, Hamlet takes quite a while over his dying. But if Claudius' killing of Hamlet lasted until Hamlet was dead, then Claudius was still murdering Hamlet after he, Claudius, was dead, since Claudius died first. Yet Hamlet was killed, and by Claudius - ignore the fact that Claudius enlisted Laertes as a willing tool, ignore it because precisely the same problem arises, Laertes dying before Hamlet. Such examples show that we cannot maintain both of two cherished principles in all cases, even if they hold in most cases: the principle that an action lasts until its primary intention is achieved (or not achieved, as the case may be), and the principle that actions do not outlive their agents. But common usage has not needed to choose between these principles, either way Claudius gets the blame, and so the concept manifest in usage is vague on this very point. Now this vagueness is unlike the last because any two candidates do not massively overlap, certain pairs are even disjoint. However, faced with such vagueness, we can still claim that actions are physical events because the vagueness is in principle dispensable: if necessary we could introduce a precise set of concepts - action\(^1\), action\(^2\), action\(^3\) and action\(^4\) - taking the old concept of action to be the genus of which these are the species. It remains true that in all cases actions are physical events.
However, PDI must be relaxed as illustrated to accommodate vagueness only if there is some argument to show that the vagueness in question is tolerable and does not demolish the identity claim. The distinction between tolerable vagueness and a vagueness which is catastrophic, in that it demolishes identity claims, is itself vague. But it is clear, I think, that vagueness puts pressure on identity claims and needs to be positively shown to be harmless, either because it is marginal as in the first example, or because it is remedial as in the second, or in some other way.

To see that anomalous monism does indeed suffer from vagueness, threatening the identity claim, consider the Officer of the Watch who forms the belief that the Bismark is approaching because he sees it approaching and who consequently telephones the Captain. Which physical event is his belief that the Bismark is approaching? All we know so far is that it was caused by a certain sensory stimulation and causes his bodily motion of lifting the telephone. We also know that between that sensory input and that motor output lies a vast causal nexus, and we know that some portions of that causal nexus are to be identified with other relevant beliefs and desires. But that is all we have to go on: no lateral comparisons with other agents or with this agent at other times are relevant to determining which physical event is the belief. For there are no psycho-physical laws, and any de facto generalisations cannot be projected from case to case: we must, so to speak, determine first the identities and then such psycho-
physical generalisations as may be. Given all the sensory input, behavioural output, other beliefs and desires which psychological theory attributes to the officer as causes or as effects of his belief that the Bismark is approaching, and given the vast causal nexus between stimulation and output, token-token identity claims are catastrophically underdetermined. It is like having too few equations to solve and too many variables to determine.

It might seem that we could do better if we considered a larger portion of the officer's psychological history. After all, some of the beliefs relevant to his telephone lifting had been acquired when he came on watch and read the standing orders of the day, and others were acquired at training sessions on recognising enemy warships. But, given that the identities we are seeking are token-token and not law-like, and given, per impossible, I claim, that we have identified those neural states at his earlier briefing and earlier training which were those relevant beliefs, and supposing that those physical states then caused persist in his head now, at the time when he forms the belief that the Bismark is approaching; even given all this, we have no warrant to suppose that those persisting physical states are still tokens of those beliefs (about what to do if you see an enemy warship, about what the Bismark looks like) when tokens of those beliefs are again causally active in his behaviour. The token beliefs caused at the briefing and at the training are of the same mental type as the token beliefs now causing his behaviour, but of course it does not follow, for Davidson, that the present
beliefs are identical to physical states of the same physical type as before. Blood flows, cells blink on and off: all we require, of whatever token physical states are currently those token background beliefs, is that they are effects of the briefing and training, and they are among the causes of his current behaviour. Every token belief of the same psychological type which is causally active at one time or another could be a token state of a different physical kind on each occasion.

That is what makes the vagueness catastrophic to token-token identity claims - almost anything would do. In contrast, the remedial vagueness we found in our concept of action was a type-type vagueness. Having tightened up our concept, having selected say action\(^3\), all token-token identities between actions\(^3\) and physical events are thereby determined. But the present case is in principle irremediable, for resolving the vagueness in one token case leaves the rest as undetermined as ever.

Davidson obscures the problem because because he occasionally lapses into permitting psycho-physical generalisations to determine psycho-physical identities, as for example in the following passage\(^{17}\):

... if an event of a certain mental sort has usually been accompanied by an event of a certain physical sort, this is often a good reason to expect other cases to follow suit roughly in proportion. The generalisations that embody such practical wisdom are assumed to be only roughly true, or they are explicitly stated in probabilistic terms, or they are insulated from counterexample by generous escape clauses. Their importance lies mainly in the support they lend singular causal claims and related explanations of particular events.
Thus, it seems, Davidson permits psycho-physical generalisations to be projected, to guide judgements of psycho-physical identity. Such a blurring of the distinction between strict projectible laws on the one hand and mere matter-of-fact generalisations on the other is a lapse, because it would undermine the argument for monism. Davidson argued that causality presupposes projectible laws. To the extent that psycho-physical generalisations are projectible and law-like they will support a relation of psycho-physical causality! We have not monism, but interactive dualism.

Davidson seeks and finds a common sortal, events, to cover the claimed identity: mental events are the same events as physical events. And Davidson gives a criterion for identity between events:

$$(e_1)(e_2)(e_1 = e_2 \leftrightarrow (e_3)((e_3 \text{ causes } e_1 \rightarrow e_3 \text{ causes } e_2)$$

& $$(e_1 \text{ causes } e_3 \rightarrow e_2 \text{ causes } e_3))$$

Unfortunately this criterion does not help at all with our problem. Suppose we know the total physical history of the officer during the relevant period of time. And suppose we have a complete psychological account of all the mental states causally connected to his belief that the Bismark was approaching. Our criterion of event identity offers no unique solution to the mental-physical identity problem.

Let me relegate to Appendix A to this chapter a discussion of Field's semantics for catastrophically vague terms, and to Appendix B Kim's discussion of supervenience, for while both show initial promise neither, in my view, provides a way to reconcile anomalous with monism.
Section (iv): Points of agreement and disagreement with Anomalous Monism summarised

Let me finally compare and contrast my answer to the question "What is the relation of cognitive states to physical states?" with that of anomalous monism, to show in particular how my answer stays clear of catastrophic vagueness.

Firstly, my answer adopts the following theses from anomalous monism:

(i) Causal relations require strict causal laws: strict in the sense that any boundary conditions can, in principle, be discharged yielding unrestricted quantifiers.

(ii) Psychological laws, in common with any second-order functional laws, are not strict in this sense.

(iii) Psychological explanations, like any functional explanations, are holistic.

(iv) Psychological facts and concepts, being normative, are not reducible to physical facts and concepts.

(v) Monism: psychological states and events are physical states and events.

However, my answer departs from anomalous monism at the following points:

(i) Psychological states and events are second-order functional states and events, constituted by first-order states and events, which first-order states and events are covered by the causal laws.

(ii) This psycho-physical identity is warranted by psycho-physical correlations between the second-order states of the psychological theory and the second-order states of
the dispositional theory, a correlation whereby the beliefs, desires, etc. attributed to the subject by the psychological theory can be identified with the history of dispositional states attributed by the physical functional theory, to within the limits of a tolerable vagueness.

(iii) Since we do not have, nor look likely to have in the foreseeable future, such a physical dispositional theory for humans, my warrant for (ii) is merely my metaphysical predudices for naturalism and against Cartesian dualism.

(iv) These second-order dispositional-cum-psychological states are multiply constituted by first-order physical states. The catastrophic vagueness which beset the identity claims of anomalous monism becomes the harmless many-one relation of constitution.

(v) Psychological states and events are not part of the causal nexus, which is confined to the first-order states which constitute them. But this permits my psychological states to be as active in the world as any other second-order state active through the causal agency of the first-order states which constitute it.
Appendix A: Field's semantics for catastrophically vague terms

Benacerraf argues that the natural numbers cannot be identified with sets because of catastrophic vagueness: given one such identification indefinitely many rival identifications may be proposed. In response to Benacerraf, and to Quine's argument for the inscrutability of reference, Field develops a semantics of "partial reference" for numerals. In effect Field provides a semantics for catastrophically vague terms. In brief, Field proposes that the numeral "5", say, partially refers to each of the fifth members of all w-sequences. Thus "5" partially refers to the set of all sets having just five members (Russell), and partially refers to Zermelo's 5, and partially refers to von Neumann's 5, and also to the fifth member of each other w-sequence. And similarly, of course, for all the other numerals. Field proposes a selection function S which, given a numeral, say "5" again, and given one partial referent of "5", selects the 'appropriate' partial referents for all the other numerals. In effect, S recovers, from all the partial referents of all the numerals, the w-progressions. Given a numeral and one of its partial referents, take any other numeral, S will select the 'appropriate' partial referent of that second numeral.

An arithmetic sentence is then defined as being true iff it is satisfied by each sequence of partial referents selected by S. For example, "7+4=11" is true iff it is satisfied by the sequence of Russelhian partial referents of the numerals, and by the sequence of Zermelo partial referents, and by ... Given similarly diverse interpretation
of the operational terms "+', "-", etc., the truth values of arithmetic sentences come out as required. In the manner indicated, Field has provided a semantics whereby catastrophically vague sentences can retain determinate truth values.

Has Field shown how numbers can be sets since the sentence "Numbers are sets" comes out true in his semantics without there being any answer to the question "Which set is the number 5?" and its ilk? If so, Field could be taken as providing a model for claiming that mental events are physical events, without there being any answer to the question: Which physical event is my belief that the moon is made of green cheese? Is even extreme vagueness not catastrophic for identity claims?

Field's semantics of partial reference are intended to explicate Quine's claim that reference is inscrutable, the inscrutable referent of "5" being replaced by all its scrutable partial referents. And Davidson is sympathetic to Quine's arguments and draws the conclusion that sub-sentential referential relations are indeterminate and have only an instrumental ontological status in explaining the truth conditions of whole sentences. That is to say: the referential relations between parts of sentences and the world are merely whatever the truth conditions of the whole sentences require them to be, and the truth conditions of whole sentences fail to determine any unique set of such relations. So Davidson may well be sympathetic to a proposal to replace the relation of reference by the relations of partial reference, and willing to have his claim of monism, and of identity between mental and physical events, interpreted in the manner of Field.
If so, we can use Field’s semantic apparatus to make both the following come out true, using "m" and "p" as variables for mental events and physical events respectively:

\[(m)(\exists p)(p=m)\]
\[(\exists p)(p=b)\]

where b is my belief that the moon is made of green cheese. The truth of these sentences would be the claim of monism, even though there is no answer to the question: Which physical event is b? The idea would be to give "(\exists p)(p=b)" the truth value of the (perhaps infinite) disjunction

\[p_1=b \lor p_2=b \lor \ldots \lor p_i=b \lor \ldots\]

where \(p_1, p_2, \ldots, p_i, \ldots\) are all physical events. That disjunction is true, by Field’s semantics, since whatever partial referent of "b" we select one or the other of the disjuncts will be true, and so the disjunction is satisfied by all "b"'s partial referents. But no singular sentence of the form

\[p=b\]

is true, since "b" has many partial referents only one of which can be p, so "p=b" is not satisfied by all the partial referents, and so is not true. (Obviously, in such a semantics, "v" is not truth conditional: the truth value of the disjunction is not a function of the truth value of the disjuncts.)

However, Field does not provide a mechanism for reconciling monism with seemingly catastrophic vagueness. It is one thing to provide a semantics whereby the sentences

\[(m)(\exists p)(m=p)\]
\[(\exists p)(b=p)\]
come out true, but another to show that those sentences still express the metaphysical thesis of monism. These sentences no longer express monism just because the semantics does give an answer to the question: Which physical event is b? That answer is: None! For every singular sentence of the form "p=b" is false according to those semantics. For "p" names, ex hypothesi, a physical event, and "b" has, ex hypothesi, many partial referents of which p is only one. So the sentence "p=b" is not satisfied by all the partial referents of "b", and so the sentence is false. The semantic theory does not show how mental events can be identified with physical events without each particular mental event being identical with one and only one physical event, modulo a tolerable vagueness.

Having thus rejected Field's semantics as a mechanism for reconciling identity with seemingly catastrophic vagueness, we are not thereby committed, and nor therefore is Davidson as interpreted by me so committed, to supposing that the referents of singular mental and physical terms are scrutable. We can still suppose that each singular term has not one referent but many partial referents, consistently with monism, provided we do so for mental and physical terms alike. Monism then implies that some singular sentence of the form "p=b" is true because "p" and "b" each have the same set of partial referents, and S selects the same member of their class of partial referents for each name. However, these observations are irrelevant to the main business in hand: which was to see whether Field's semantics could counter the threat to monism presented by the vagueness of "b", etc.
relative to "p", etc., the vagueness of mental referents relative to physical. To this issue it is irrelevant that a mutual and balanced vagueness is no threat to monism.
Appendix B: The supervenience of the psychological upon the physical

Perhaps a different tack will counter the treat of catastrophic vagueness. Perhaps we should invoke the supervenience of the mental upon the physical. Davidson claims

> Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.

Such supervenience would permit comparisons with other cases to constrain our token-token identifications, and thus supervenience holds out the hope that by increasing the number of available 'equations' we could achieve unique 'solutions', or settle for a tolerable vagueness. For supervenience to play this role it must be projectible, and indeed we note Davidson's modal language: "cannot be" and "cannot alter".

My predilection for naturalism commits me to such a supervenience: How could physical facts determine psychological facts (although without psychological facts or concepts being reducible to physical facts or concepts) if two entities could have all their relevant physical facts in common, and yet differ in their psychological facts. Indeed, supervenience is all that naturalism without reduction amounts to, it would seem.

Jaegwon Kim explores this conception of supervenience,
and attempts to prove, pace Davidson, that supervenience implies psycho-physical laws. 20 However, as we shall see, the psycho-physical laws which Kim proves are no real threat to Davidson’s anomalism, but, for that very reason, they fail to reduce a catastrophic to a tolerable vagueness.

Kim defines the supervenience of M properties on N properties as follows 21:

a family M of properties is supervenient on a family N of properties with respect to a domain D of objects just in case, necessarily, objects in D which share all properties in N# will also share all properties in M#.

where N# and M# are the closures of N and M respectively under the normal Boolean operations of complementation, conjunction, disjunction, etc. Kim continues 21:

We also define a subset M* of M# as follows: if M is finite, each member of M* is a maximal consistent conjunction of the properties in M, and the complements of the properties in M; if M is not finite, the members of M* are maximal consistent sets of properties in M and their complements.

Briefly, Kim then proceeds to prove, for any given property P of M# and any object x from D which instantiates P, there will be some property Q of N# such that x has Q and

\[(x)(Qx \rightarrow Px)\]

So it seems we have a psycho-physical law.

Kim achieves this result by taking Q to be that N* which characterises the given subject. In a nutshell his argument runs as follows: Given x and P from M, take that N* (a maximal consistent set of properties from N) which characterises x. Nothing that shares N* with x can differ from x in any N# property. So, by the hypothesis of supervenience, anything that has N* also has P of M.

We invoked supervenience to try to dispel catastrophic
vagueness, and indeed that was one of the purposes Kim himself had in mind. But unfortunately the laws which he has shown to be guaranteed by supervenience are not likely to further our project. Consider first a community no two members of which have exactly the same total psychological history - likely ours is such a community. It follows, by the supervenience hypothesis, that no two members have the same total physical history - where the physical properties in question are those upon which the psychological properties are supervenient. (The same functional theory may apply to a, b, c, ... and yet a, b, c, ... be in different concrete functional states.) In such a community supervenience is trivially satisfied, because no two members are in the same total physical condition. And if a is in the total physical condition $Q^a$, and in the psychological states $P^a_1, P^a_2, ..., P^a_i, ...$ the relevant Kim-type laws will be:

\[(x)(Q^a_x \rightarrow P^a_1x)\]
\[(x)(Q^a_x \rightarrow P^a_2x)\]
\[
\vdots
\]
\[(x)(Q^a_x \rightarrow P^a_i)\]
\[
\vdots
\]

None of these will help us identify $P^a_1$, say, with a particular proper part of the total state $Q^a$. And further, comparison with other members of the community will not help, because their Kim-laws have different antecedents, e.g.

\[(x)(Q^b_x \rightarrow P^b_1x)\]
\[(x)(Q^b_x \rightarrow P^b_2x)\]

etc.

The laws do not provide any principled way of relating
particular proper parts of a subject's total (psychologically relevant) physical state to any of his particular psychological states.

Suppose now that some a, b, c, ... do coincidentally share the same total physical history $Q^a, b, c, \ldots$ and hence the same total psychological history $P_1, P_2, \ldots, P_i, \ldots$. All we are guaranteed is a set of laws of the form 

$$(x)(Q^a, b, c, \ldots x \Rightarrow P_i)$$

but, again, no clue how to uniquely pick out some proper part of the total state to identify with $P_i$, and another, different proper part of the same total physical state to identify with $P_j$, etc.

Thus Kim's psycho-physical laws are not of a kind to which Davidson need object. But for that very reason supervenience is too weak a constraint to generate a unique, or near unique, mapping from particular parts of his psychological history onto particular parts of a subject's (psychologically relevant) physical history. Anomalous monism remains beset by catastrophic vagueness.

In my view, not only are psychological facts supervenient upon the second-order physical facts which realise them, but also upon the first-order physical facts which constitute the second-order facts. A machine cannot alter its machine state without a first-order change of state; two machines cannot be alike in all their first-order states and yet differ in their machine states.
Footnotes to Chapter One
1 Hume famously declared that one cannot derive an "ought" from an "is", and in modern philosophy I am following the lead of Wittgenstein (1) and the discussions of Kripke (3), McGinn (4) and Wright (1).
2 Notably by Fodor (2) and McGinn "The Structure of Content" in Woodfield (1).
3 Of course, I mean that causal relations between first-order states or events are logically contingent, not causally contingent. Some may dispute this if, like Davidson "The Individuation of Events" (in Davidson(1)), they individuate events by their causes and effects. In reply I would plead that I am using the terms "event" and "state" loosely, as placeholders for whatever physical entities they do admit, between which the relevant contingent causal relations hold. And surely everyone admits that, at some level of analysis, the causal inter-relations which constitute a contingent mechanism are themselves contingent inter-relations.
4 Putnam: "Minds and Machines" in Putnam (1)
5 Putnam: "Minds and Machines", "Philosophy and our Mental Life", "The Nature of Mental States", in Putnam (1)
6 It follows that second-order functional organisation cannot be the real essence of a natural kind. If tigers, humans, etc. are natural kinds, and it is of their essences that they realise certain second-order functional theories (physiological and psychological), then that can only be part of their essences, at most. It must also be part of their essence how those functional theories are to be constituted by first-order states.
7 To say "we can refine causal laws until any ceteris paribus clauses are discharged" is to speak loosely. If we take laws to be man-made entities which constitute theories, then there is no guarantee that such laws will be achieved. Davidson's claim would be that it is the aim of the natural sciences to achieve such laws, science progresses by finding laws ever closer to that goal, and
there is no bar in principle to that progress being maintained. On the other hand, if we understand laws of nature realistically, to be the real rules according to which the world operates, then Davidson's claim would be that those rules, whether we can discover them or not, have the features described. (I ignore the irrelevant complication that the 'ultimate' laws of nature may be statistical.)

8 Davidson: "Mental Events", in Davidson (2) p 213/4
9 Ibid. p224
10 Ibid. p222

11 Geach has championed a concept of relative identity, in Geach (1) chapters 7.1 and 7.2, such that the following could both be true: a is the same mental event as b, a is not the same physical event as b. On such a view there is no answer to the question: Which physical event is the psychological event a? For it is both a and b, two different physical events. However, I cannot make sense of such a notion of relative identity, and would deal with the examples which superficially seem to give it support in the manner of Wiggins (1).

12 The philosophers I have in mind are, in order, Hornsby(1), McGinn (2), Davidson (2) and Thalberg (1).

13 Benacerraf (1). Benacerraf goes on to argue that numbers are not objects at all, but I ignore that part of his argument.

14 Field (1)
15 Davidson:"Radical Interpretation " in Davidson (3)
16 Davidson:"Reality without Reference" in Davidson (3)
17 Davidson (2) p218/9
18 Davidson:"The Individuation of Events" in Davidson (2)
19 Davidson (2) p214
20 Kim (1) and (2)
21 Kim (1) p152
22 Particularly in Kim (2)
Chapter Two: Conceptual Content and Psychological Realism

Section (1): Introduction

In chapter one perceptions, beliefs, desires and intentions were taken to be real states and events, identical with second-order dispositional states. It may seem therefore that I am in no position to question the reality of psychological facts: a psychological fact being simply a dispositional state or event instantiating a psychological concept. Suppose Tom believes that the moon is made of green cheese. It is open to question whether the dispositional state of Tom which is that belief really has the conceptual content and therefore the semantic content of the sentence "The moon is made of green cheese". I.e. it is open to question whether Tom's beliefs, desires etc. really are representational states. In this chapter I shall argue that they are in the case of Tom and his fellow language users.

Loar1 develops as interesting account of the role of content clauses in propositional attitude ascriptions used to explain behaviour. He claims that the sentence governed by the psychological verb is an index of the psychological state reported by the sentence as a whole. The index is not merely the sentence qua syntactic structure, but the sentence qua syntactic-structure-plus-interpretation (fine-grained truth conditions, thinks Loar2). Loar thinks the psychological verbs (perceives, believes, etc.) classify internal states or events by their board functional roles, and then the content sentences index the precise
functional roles. For a sentence $s_1$ to index a psychological state, say a belief, is, briefly and roughly, for that sentence to be one of a set of sentences $s_1, s_2, \ldots, s_i, \ldots$ whose inferential inter-relations mirror the functional inter-relations of the belief-states reported by respectively $\exists z \text{ believes } s_1$, $\exists z \text{ believes } s_2$, $\ldots$, $\exists z \text{ believes } s_i$, $\ldots$. And we will need to include perceptions, desires, etc. in the network. This rough sketch of Loar's account will suffice for present purposes.

To treat content sentences in propositional attitude reports as indices of the attitudes reported is, of course, to be realistic about those attitudes: they are taken to be real states and events whose real inter-relations are mirrored by the inferential relations between the indices. And further, regarding content sentences as indices of physical states, is consistent with the normative concepts which describe the inferential relations between the indices being irreducible to physicalistic concepts, even to whatever physicalistic concepts describe the inter-relations between the states so indexed.

Taking content clauses to be indices of the psychological states reported leaves open the question: Do the psychological states so indexed really have the conceptual (and semantic) contents of their sentential indices, or are those conceptual (and semantic) contents to be treated instrumentally? The necessary and sufficient condition of a system of indices is that the relations invoked between the indices mirror the relevant relations between the states indexed.
Any system which achieves this end is an index of those states. Loar recognises this point: if we had an indexing system which consisted of context-free interpreted sentences inter-related by 'topic neutral' syntactically defined inference rules and 'topic specific' syntactically defined meaning rules, then an equally good indexing system could be constructed employing the Gödel numbers of those sentences and arithmetic functions of those numbers in place of the inference rules.

Now there are obvious reasons for preferring our familiar indexing system to such a Gödelian alternative: I may learn that \( p \) on the ground that \( \neg p \) indexes Tom's belief state, he being an expert well-motivated to find out whether \( p \); or, on learning that Tom studies classical music I conclude, without further ado, that he will be more familiar than I am with the major works of Bach and Handel. On the other hand, there are no inferences from the Gödel numbers indexing Tom's beliefs to how things are in the outer world, nor from how things are in the outer world to Tom's Gödel indices, assuming, of course, that the Gödel numbers are being applied to Tom's states directly, and not as a secondary indexing system computed from the sentential indices. True enough, also, that the familiar indexing system is graspable whilst its Gödelian alternative probably is not, because the Gödel numbers and their arithmetic functions will be a massively ad hoc lot. But again, this pragmatic difference does not make the one an indexing system and the other not. So far as the job of recording
Tom's transitions from one internal state to another goes, either indexing system does the job.

Given such a numerical indexing system of psychological states, we do not suppose that the states themselves are numerically inter-related; we do not suppose that, if the index of one state is the cube of the index of another state, then the one state indexed is the cube of the other. So using a sentential indexing system does not commit Loar to supposing that the states so indexed have the conceptual contents of their indices - nor, therefore, the semantic contents of those indices, taking their conceptual contents plus contexts to determine their semantic contents. What is to be interpreted realistically is only that the states and events indexed are inter-related by some system of relations or other which mirrors the normative relations of the sentential indices, and also mirrors the arithmetic relations of their Gödel numbers. It is left an open question whether the inter-relations between the states indexed are conceptual and semantic relations; i.e., whether the states indexed are themselves representational.

However, given my prejudice for naturalism, I believe that physical facts do determine (non-reductively) such non-natural facts as the conceptual contents of sentences. An attractive picture is that the first and second-order physical facts which constitute and realise the inter-related psychological states and events, determine conceptual contents for those psychological states and events, the same conceptual contents as are found in their sentential indices.
Indeed, this picture would locate which physical facts determine the conceptual contents of sentences - viz. those physical facts which determine the conceptual contents of the psychological states which those sentences index. An alternative picture, but still a naturalistic picture, is that other physical facts, perhaps facts about the public contexts of utterances, determine the conceptual contents of sentences, and the psychological states indexed by those sentences do not really have conceptual contents. However, this alternative picture is less attractive because it makes a mystery of how we grasp the conceptual contents of sentences. And since language is our artefact it would be a mystery if sentences had conceptual contents which we were unable to grasp.

So, I claim, if God created physical facts according to the Gõdel-number blueprint, he would thereby have created states and events with the conceptual contents of their sentential indices. But such a claim goes beyond seeing the content clauses of psychological reports as mere indices of internal states, and requires some positive argument for its support. In this chapter I shall assume that the sentential indices have determinate conceptual contents (an assumption to be defended in chapter three) and argue that the conceptual contents of the sentences of our home language, by which we psychological subjects index our own psychological states, those conceptual contents really do apply to the psychological states so indexed.
Section (ii): A criterion for psychological realism proposed

In this section I shall propose a criterion for deciding whether or not the states indexed really do have the conceptual contents of their indices. Let us call the view that the states indexed have the conceptual contents of their indices psychological realism. And let us call the sequence of a person's conceptual indices over time his psychological history. My criterion is, then, a criterion for psychological realism, for taking a person's psychological history realistically.

The basic idea is this: To be interpreted realistically, the psychological history of Tom, say, must show that there was something it was like to be (cognitively and volitionally) Tom to think Tom's thoughts. I take this to imply that there is not a second application of the psychological theory to the same set of Tom's internal states and events, which assigns indices with different conceptual contents, but which maps the inter-relations between those internal states of Tom equally well. If a second application of the sentential indexing system records the same naturalistic history, the same sequence of transitions from internal physical state to internal physical state, but yields a different crop of sentential indices, different in their conceptual contents, then each indexing defeats the claim of the conceptual content of the other to be taken realistically; since Tom cannot really be in two minds at once. Moreover, this necessary condition also seems sufficient: I cannot see what more could be required for taking the conceptual
content of the psychological history realistically.

If the physical states of Tom which are being indexed were first-order states, then we might expect that under one indexing the states \( s_1 \) and \( s_2 \) were indexed by the sentences \( S_1 \) and \( S_2 \) respectively, and under a second indexing \( s_1 \) and \( s_2 \) were indexed by \( S_2 \) and \( S_1 \) respectively, and yet both indexings equally met all reasonable constraints on psychological theorising. However, the states being indexed are second-order. So we can make such permutations of the indices which do not change the overall psychological history of the subject impossible in principle: we can admit that the same second-order physical states, indexed by \( S_1 \) and \( S_2 \) respectively, are constituted by the first-order physical states \( s_1 \) and \( s_2 \) respectively, and are also constituted by the first-order physical states \( s_2 \) and \( s_1 \) respectively.

We hope that the reality of Tom's psychological history shows what it is for Tom to grasp the conceptual contents of those indices. So any other system of sentential indices, whose inferential inter-relations exactly match the dispositional inter-relations of the states indexed, expresses the same conceptual contents.

Our ability to construct a psychological history of Tom is dependent upon the repertoire of conceptual contents which make up our own, the reporter's, psychological lives. May the uniqueness of a system of indices not be due merely to the conceptual paucity of the indexing language? Perhaps adding a richer range of conceptions to the indexing
language would generate a rival set of indices, thus destroying the claim of the original set of conceptual contents to be taken realistically. But whether or not there really is something it is like to be Tom, to think Tom's thoughts, cannot be an artefact of the conceptual richness or paucity of the language we employ to theorise about Tom. Whether we can formulate a psychological history of Tom will depend upon the conceptual resources of our indexing language, but not whether Tom has a psychological life, nor what such life he has.

However, I hope to show in section (v) below that no enlargement of the conceptual resources of the language employed to index Tom's attitudes will produce the means for a second indexing of Tom's attitudes which (a) assigns different conceptual contents to those attitudes and (b) equally exactly indexes those states, provided (c) that Tom is himself a language user and the indexing language is Tom's own, and (d) the conceptual contents of the sentential indices are determinate.

Condition (a), that two systems of sentential indices assign different conceptual contents, is vague. Obviously Tom's perceptual input and motor output associated with the one system of sentential indices must be either the same as or consistent with the perceptual input and motor output associated with the other system of sentential indices - otherwise both systems of indices could not be applicable to one and the same Tom. And clearly one can be doing two things at once, thinking two thoughts at the same time - as when I whisper loudly both to greet a friend and to annoy
the visiting speaker. But I take each system of sentential indices to be intended as a **complete** psychological history of Tom during the relevant period. And reading each set of indices as a complete history of Tom, I rely on a commonsense judgement whether they are two versions of the same history or two rival histories; whether, if Tom thought the thoughts ascribed by the one, he did or did not think the thoughts ascribed by the other. Condition (b), that the two systems of indices are equally exact, is also vague. Indices mirror the functional relations between the states indexed thanks to the inferential relations between their (the indices) conceptual contents. By an *exact* indexing I mean that there are no inferential relations between the indexing concepts which do not mirror a functional realtion between the states indexed. Given a system of indices which is exact in this sense, then, no doubt, one could find a mapping of those indices onto other sentences with richer conceptual contents which preserved the inferential inter-relations of the first system, but where the inferential inter-relations of the image sentences were richer and under-employed as indices.

Of course, the psychological history of Tom to date, generated by applying psychological theory to Tom to date, is defeasible: it may be revised in the light of further evidence. And so the known psychological history of Tom to date is not unique, a later conquering rival psychological theory will, in general, fit the earlier evidence equally well. But this epistemological defeasibility, and consequent non-uniqueness of a psychological history, is not
to the ontological point: that point being that the totality of Tom's physical facts cannot determine two sets of incompatible psychological facts. But to avoid the perhaps illegitimate notion of the totality of physical facts about Tom, my criterion for psychological realism should rather be: For any given pair of psychological histories (= sets of sentences) assigning different conceptual contents to Tom's thoughts which are equally supported by a finite corpus of physical facts, that corpus of physical facts can be enlarged to produce a set which eliminates one of those histories - if the surviving history is to be interpreted realistically.

The physical facts which constitute and realise the psychological facts about Tom are first and second-order facts about, at least in part, what goes on inside Tom's skull. But the evidence we go on when formulating a psychological history of Tom are not these facts. It is curious that the facts which are readily available to us and which warrant our indexing of Tom's internal states are not those internal states. I conjecture that it is unnecessary condition of those physical facts about Tom constituting and realising psychological facts about Tom that Tom is not aware of them as physical facts and not aware of their physical inter-relations.

Section (iii): Does narrow functionalism fail the test for psychological realism?

I doubt that narrow functionalism will pass my test for psychological realism. Adapting an example from that
arch-narrow-functionalist Fodor, suppose a government department has a computer program in which it stores and manipulates information about cities. An operator may feed in the datum sentence, in programming language

Boise is a city

but the compiler replaces this by, says Fodor

'Boise' is stored at the address CITY

in the machine language. Fodor's point is that the process of compilation is not translation, because the semantic content of the first sentence - that Boise is a city - is quite different from the semantic content of the second - that a certain piece of syntax is stored at a certain address - so the conceptual contents of the two sentences must also be quite different. Fodor's general point is that the programming language version of a program has a different conceptual content from a machine language version of the same program.

Now suppose that we have a narrow psychological theory of the behaviour of my dog which attributes to it beliefs, desires, etc. about rabbits, bones, etc. And suppose that a cognitive psychologist produces a 'machine code' account of those same inner-state-to-inner-state transitions. But where my psychological theory attributes to a given inner state the conceptual content of desiring to kill a rabbit, the cognitive psychologist attributes to that same inner state the conceptual content of desiring to store the symbol 'Rabbit' at the address DEAD. My psychological history of my dog indexes its internal states with concepts of bones,
rabbits, etc. whilst the cognitive psychologist indexes those same states with concepts of syntax and addresses. And no doubt a third, Quinean, cognitive psychologist could index those same inner states with conceptions of preferred patterns of surface irritations, nerve hits.

The animal's behaviour is as normatively rational under one psychology as under the others, at least a narrow functionalist must suppose so. Under one it desires to kill rabbits, under another to store as many symbols 'Rabbit' at the address DEAD as possible, under the third to achieve certain patterns of surface irritation, and its day will have been as successful or as disappointing under one as under the others. Further, it is no objection to argue that the dog cannot believe that 'Rabbit' is stored at the address PRESENT because the dog cannot perceive its inner states. True enough, when we perceive a rabbit we come to believe that a rabbit is present due, in part, to a causal chain stretching from the rabbit to ourselves. But the dog may believe that 'Rabbit' is stored at the address PRESENT simply because that belief is constituted by 'Rabbit' being stored at the address PRESENT, no mediating process being required.

It seems, then, that by my criterion the dog's inner states have no conceptual contents, that the three theories are to be interpreted instrumentally. And, of course, what goes for the dog presumably goes for us humans too. Presumably there is a 'machine language' indexing of our inner states, narrowly construed, the same inner states
that commonsense indexes with concepts of cabbages, kings, chewing gum and string.

The moral of this tale is, to my mind, wide functionalism. I shall argue in chapters four and five that the first and second-order physical systems whose states constitute and realise Tom's psychological states extend beyond his bodily envelope, to include his normal perceptual and manipulative interactions with his physical environment. If this is indeed so, then the second-order states which are Tom's beliefs, desires, etc. are not identical with the second-order states which the solipsistic cognitive psychologist indexes with his machine language sentences. For the functional system which defines the first set of second-order states extends beyond the bodily envelope whilst the functional system which defines the second set of second-order states is confined within the bodily envelope. Of course, we expect an interesting relation between the two functional systems: the first-order states which constitute the narrow second-order states will be proper parts of the first-order states which constitute the second-order states of the wide functional system, and this in some systematic way such that the conceptual content of the narrow second-order state is the compiled version of the conceptual content of the wide second-order state. Or rather, so we may hope.
Section (iv): The psychological status of dumb animals

My criterion for applying the conceptual contents of indices to the internal states indexed presents two problems: firstly, what counts as one and what as two psychological histories, and secondly, how are we to establish a negative, that there isn't a second psychological history (= set of sentences) of the creature in question? In the case of the psychological histories we devise for dumb animals I strongly suspect that we cannot answer these questions and I am doubtful whether those histories should be interpreted realistically. To interpret the psychological histories we ascribe to dumb animals instrumentally is merely to exercise due epistemic caution. It is not to deny that animals have a cognitive life, grasp and exercise concepts, since I cannot positively provide alternative systems of indices which ascribe different conceptual contents to the animals' functional states and mirror those states with equal exactitude. It is merely to refrain from ascribing the conceptual contents of the indices which we do employ in explaining their conduct to the states so indexed.

In the case of fellow language users, I hope to show in the next section, we have grounds for optimism, because fellow language users are self-indexers of their own states - remember, I am assuming in this chapter that those indices have determinate conceptual contents. But first, our dumb friends.

Undoubtedly we are anthropomorphic in our practice when we ascribe cognitive and volitional states to dumb
animals. We say that the dog is digging for the rabbit it saw disappear down the burrow. But do we really suppose that the dog knows or cares whether it finds the same rabbit as it saw disappear? Wouldn't any rabbit do? If so, then we should say that the dog's hatred is directed at rabbit features rather than rabbits, it having learnt how one rabbit feature can lead to another.

Peacocke tackles this problem and seeks to formulate what is in effect a necessary condition for psychological realism: that condition he calls his tightness constraint. Roughly speaking, if A and B are two psychological histories (sets of sentences) which both fit the creature's behaviour, and C(A) and C(B) are the sets of concepts which each history attributes to that creature, then A is tighter than B iff

(i) Some condition is frameable using C(B) but not C(A), and no action in the creature's repertoire is explained in B using this condition, and

(ii) the reverse does not hold.

We may not interpret the history B realistically if there is another history A which is tighter than B. Of course, Peacocke realises that the tightness ordering may only be partial. I have no quarrel with Peacocke's tightness constraint, but an interesting feature is that Peacocke sees his tightness constraint as a generalisation of a constraint on radical translation elucidated by Evans, thus linking linguistic and non-linguistic behaviours.

Evans sees that, given determinate translations of the
Native's signs for assent and dissent, and given determinate translations of Native observation predicates like "white", as Quine permits, we can disambiguate the translation of "gavagai" as between "rabbit" and "undetached rabbit part" by seeing whether natives assent to "White gavagai" in the presence of a white undetached tail of an otherwise brown rabbit, and we can disambiguate the translation of "gavagai" between "rabbit" and "rabbit feature" by seeing whether the natives assent to "White gavagai" in the presence of a white rabbit-shape formed by the juxtaposed white tails of a performing troupe of otherwise brown rabbits. Evans general point is that an acceptable translation manual must permit the assertion conditions of compound sentences to be uniformly generated from the meanings of their component parts. And the indeterminacy Quine claims for the radical translation of "gavagai" could be resolved by applying this compositional constraint. Peacocke connects Evans' rebuttal of Quine with his own tightness constraint in the following passage:

Very roughly, the idea is that just as the ascription of semantical properties [to sentences] is answerable to the existence of certain kinds of sentences, in a corresponding way the attribution of concepts is answerable to the existence of certain kinds of action. The generalisation would run: we should not construe someone as having propositional attitudes, the content of which is specified by using the concepts of objects of a given kind with certain boundaries, if there is a tighter set of attitudes we can ascribe which does not involve attributing a concept determining objects with the same boundaries. Evans' requirement is indeed a special case of this requirement, restricted to the semantical, since saying something or assenting to something is an intentional action.

Let me confess, first of all, that I cannot see Evans' rebuttal of Quine as a specific instance of Peacocke's
more general tightness constraint. The tightness constraint seems aimed at eliminating a conceptual repertoire which permits distinctions, generates truth conditions, which the creature simply does not manifest any grasp of in its behaviour. That is to say, its behaviour neither manifests a content that p nor a content that not-p, where "p" expresses this excess conceptual content. This understanding of his tightness constraint seems clear from both his formulation of it and from the bulk of his discussion concerning it. But Evans' rebuttal of Quine does not show that the conceptual repertoire of the victorious translation manual fits overall native usage more tightly than its rivals. Indeed the overall conceptual repertoires of the rival manuals may be exactly the same: C(A) = C(B) = ...

Native usage, if Evans is right, shows that the rejected translations were merely wrong, attributing false psychological histories to the native. But the verdict might have gone the other way: it was equally within the normal repertoire of native action to manifest the concept of an undetached rabbit part by their utterances of "gavagai". (Consequently, given that the correct translation of "gavagai" is "rabbit", it will come as no surprise that the natives also have the concept of an undetached rabbit part, which they manifest by uttering "gavagai-on-bit".) In contrast to this, dog behaviour manifests an interest in rabbit features, but a new kind of behaviour would be needed to manifest an interest in rabbits as distinct from rabbit features.

Thus rival translation schemes ("gavagai" - rabbit, "gavagai" - undetached rabbit part, etc.) have not been
shown to credit the native with **different** conceptual repertoires, but merely to assign the same stock of concepts to different words and phrases of Native. So resolving that indeterminacy of assignment by Evans' technique is **not** an application of Peacocke's tightness constraint.

However, whether or not Peacocke has correctly identified the relation between his tightness constraint and Evans' rebuttal of Quine, is incidental to the main point I now wish to make: the degree of freedom permitted the radical translator is **an order of magnitude less than** the degree of freedom permitted by the tightness constraint when the latter is applied to the behaviour of dumb animals. On the one hand, the radical translator has the **global** problem of translating a word type "gavagai", say, where his translation is constrained by the resulting concomitant translations of all the sentences containing that word, and, further, the syntax of Native is assumed to be determinate and known, such that the relevant corpus of inter-related sentences containing "gavagai" is **identifiable independently of** the rival translations of "gavagai". On the other hand, just these two features are missing from the application of the tightness constraint to the behaviour of dumb animals. What concepts we see manifested in a creature's behaviour depends upon how it does behave on this occasion and on other **relevant** occasions, and upon how it would behave and would have behaved in **relevant** counterfactual circumstances. But we cannot sort its pieces of behaviour into **relevantly related groups**
prior to and independently of the intentional content we take each token piece of behaviour to have. The degree of freedom is clearly different in the two cases: two systems of concepts may fit the dumb animals's behaviour with equal snugness, for all Evans' rebuttal of Quine has shown.

Of course, Quine and Evans have idealised the predicament of the radical translator, legitimately for their purposes: one can't in practice and couldn't in theory entirely settle the syntax of Native prior to its interpretation. However, I do not see how a more realistic account of translation would substantially weaken the contrast. Although, in practice, the radical translator's judgements as to the syntactic structure of Native are revisable in the light of the emerging translation manual and native behaviour, at any stage the radical translator's task is the global one of translating a (supposed) word or phrase type. And the grouping of utterance tokens by type, upon which types his analytical hypotheses operate at any stage in his investigations, must, by and large, be a grouping by natural behavioural type: utterances of tokens of the same word type must be utterances of tokens of the same noise type, more or less. On the other hand, applying the tightness constraint to animal behaviour, carries no parallel presupposition that a taxonomy of behaviour types yields a taxonomy of action types. It is notorious that tokens of the same non-verbal behaviour-type can be tokens of quite disparate action types.
In short, the conceptual content assigned an utterance token is constrained by the contents assigned to all other tokens of that type, where tokens of the same word-type are tokens of the same physical type. Of course the conceptual content assigned to tokens of non-verbal behaviour are constrained by the conceptual contents assigned to other tokens of non-verbal behaviour, but there is no presumption that the conceptually relevant other behaviour tokens are of the same physical type. The two problems, assigning conceptual content to linguistic and to non-linguistic behaviour, are of different orders of difficulty.

In the linguistic case, there is even a presumption, at least for the individual primitive words of the language, that different words, different physical utterance types, mark different conceptual contents. As Fodor remarks, we should not expect the different primitive word-types of a natural language to express the same content. (Actually, we find just this kind of redundancy in the case of especially fascinating concepts - think of all the colloquial verbs, euphemisms and slang, for sexual intercourse or for dying.) However, this presumption cannot be extended to compound sentences, where, because of the compositional nature of meaning, we expect different sentence types to express the same conceptual content.

For the above reasons, given my conceptual resources, I have no confidence that only one psychological history (set of sentences) of my dog can be constructed, attributing one set of conceptual contents to its inner states. Should I conclude that my dog does not really have a cognitive and
volitional life? Fortunately, I need not. I can refuse to assert that there is something it is like, cognitively and volitionally, to be my dog, but this does not commit me to asserting that there is nothing it is like, cognitively and volitionally, to be my dog - as I do assert that there is nothing it is like to be, cognitively and volitionally, an amoeba or this typewriter. In general, I may refuse to assert a sentence of the form \((\exists x)(Fx)\) where I cannot put myself in a position to assert an instance of the form \(Fa\). I cannot specify what it is like to be my dog, so I can neither say that it does manifest some unique set of concepts, nor that it does not.

But perhaps a Martian could do better. Perhaps a Martian's X-ray vision enables him to read the dog's language of thought. If so, the Martian need have no greater problem ascribing thoughts to my dog than in ascribing thoughts to me: I produce external sentence tokens of English, my dog produces internal sentence tokens of MENTAL(DOGGEREL?). In ascribing content to the dog's thoughts the Martian has to interpret the relevant sentences of DOGGEREL, the interpretation of a given sentence being constrained by its vocabulary and syntax, and thereby by the correlative interpretations of all related sentences, as with public utterances.

Fodor recommends\(^{16}\) his language of thought hypothesis on two distinct grounds: firstly, as an empirical hypothesis, to explain, for example, why subjects take longer to grasp a double negative sentence than its straightforward positive,
the explanation being the extra syntactical transformation needed to process it, and secondly by challenge, asking what model other than symbol manipulation we have for computation? But, as Peacocke argues\textsuperscript{17}, these reasons do not motivate a strong language of thought hypothesis, as needed by the Martian interpreter, rather than a trivial, relational, language of thought. The strong language of thought employs first-order properties of a given representational state as its syntax, and the computational rules are defined on the first-order properties of the 'sentences', just as the syntax of a public utterance is (ideally) a first-order property of that utterance, recognisable independently of its inferential relations to other sentences. Whereas, the weak language of thought hypothesis sees the syntax of an internal state as a second-order property of it, as a matter of its causal relations with the other sentences tokens of MENTAL. Now clearly, the radical interpreter sees the syntax of public utterances as first-order properties of those utterances (though, which first-order properties of utterance are its syntax is a matter of its relations with other utterances and their syntax).

So, if the Martian is to employ the sentence tokens of MENTAL as we employ utterances when attributing psychological states, we need the strong language of thought hypothesis. However, the trivial, relational language of thought hypothesis is more plausible. According to this hypothesis, the syntax and content of a token are determined not by its first-order intrinsic properties, but by its neural connections,
its causal relations to other tokens. Consider, for example, a little mechanism for effecting the computations:

$$P \vdash P \lor Q ; \quad Q \vdash P \lor Q ; \quad -P, P \lor Q \vdash Q ; \quad -Q, P \lor Q \vdash P$$

According to the trivial, relational hypothesis we might have a network of neurons inter-connected thus:

Here the contents of the token neurons represented by circles and squares is determined by their inter-connections 1 to 15. $P$ has its content because of the connection 1, giving it input from a $P$-detector, and $Q$ similarly because 2 connects it to a $Q$-detector. Connections 6 and 7 give the content of $P \lor Q$. Clearly, if inferences of the forms $P \lor Q, -P \vdash Q$ and $P \lor Q, -Q \vdash P$

are to perform any useful service, we require that $P \lor Q$ can be activated indirectly from collateral information, since otherwise we would always be moving trivially from, say, $Q$ to $P \lor Q$, and from $P \lor Q$ and $-P$ back to $Q$. Hence the input 12. For the disjunctive syllogism we require a neuron which fires only when stimulated by both of its inputs together. Finally, there would be no point in the disjunctive syllogism if the conclusion, say $P$
merely fed back into \([\overline{P}\lor Q]\), since the latter was one of
the premisses from which \([\overline{P}]\) was derived. Hence 3 and 4.
Let me emphasise that, in the diagram, the contents
inscribed in the boxes are not first-order properties of
those neurons, but contents generated by their inter¬
relations: \([\overline{P}]\) and \([\overline{Q}]\), \([\overline{-P}]\) and \([\overline{-Q}]\), \([\overline{P}\lor Q\land -P]\) and
\([\overline{P}\lor Q\land -Q]\) respectively have identical first-order properties.
They are interchangeable parts from the body-building shelf.

The difference between the strong and this trivial form
of the language of thought hypothesis becomes evident when
we consider the operation of the computational process:
\([-Q, P\lor Q] \rightarrow P\). Let us use solid lines to indicate active causal
pathways, and dotted lines to indicate dormant causal

The crucial points to note are that \([\overline{P}]\) has the content it
does because of the connection 1, even though that causal
path was not active on this occasion. The computational
rule employed on this occasion is defined on relations
which were not causally active on this occasion. The strong
language of thought hypothesis contrasts sharply here: the
syntactical rules employed in any computational process are
defined only on first-order features which were causally
active in the process.

I hope it is evident from this brief discussion that
the trivial relational language of thought hypothesis can
meet Fodor's empirical evidence, and his challenge for a
model for computation, as well as does the strong language
of thought hypothesis. Further, it is empirically more
plausible, more in keeping with what we do know of the
contents of skulls. And finally, MENTAL, so conceived,
differs from public language precisely in a way which makes
Evan's techniques for disambiguating public utterances
irrelevant to reading neural signals: the corpus of
utterances containing the term "gavagai" are supposed given
prior to assigning conceptual content to that term, but the
syntax of a neuron and its conceptual content are alike
determined by its relations in the neural network. Perhaps
the Martian, seeing the neural activity within the dog's
skull as well as its overt behaviour, has more chance of
finding a unique set of attitudes to attribute to the dog
then we do, who only have the overt behaviour to go on.
But, if so, it will not be, I conjecture, because he can
use the contents of the skull as we use a native's utterances
to determine the contents of his thoughts.
Section (v): The psychological status of fellow language users

Does my reader fare better than my dog? Do I have a warrant to assert that there is something it is like, cognitively and volitionally, to be one of my fellow language users? Of course, I do not ask this question in order to discover the answer: that I can ask the question confirms to me, were confirmation needed by me, that I have a cognitive and volitional life, and since I am a creature of the same physical kind as my reader, my naturalism, my view that physical facts determine psychological facts, commits me to the view that my reader too has a cognitive and volitional life. Rather than trying to construct an argument for other minds, the purpose of this section is to use the datum that we fellow language users have cognitive and volitional lives to develop my criterion for psychological realism.

My criterion of psychological realism was, briefly, that there be one psychological history of the creature; that it is of one mind; that there is something it is like, cognitively and volitionally, to be that creature. The question arose: When do two sets of propositional attitude ascriptions, with different conceptual contents, impute different proper parts of the same psychological history and when do they impute rival psychological histories? In section (ii) I supposed that each set of sentences purported to be a complete psychological history of the subject during the relevant period, and hence the two sets of sentences would be rival histories if they did not impute the same total network of conceptual contents. However,
the assumption that our psychological explanations purport to be complete is empirically false, and, given holism, unrealizable. So we need some other way of determining whether the criterion is satisfied, of determining whether two psychological histories (sets of sentences) are rivals, or are proper parts of the same total history.

With Tom, a fellow language user, we can settle the issue as follows: Is he disposed to sincerely avow both sets of attitude reports, or to avow one set and deny the other? Of course, we will need to allow, at the margins for self-deception, and for those for those things so secret Tom will never admit to. But such marginal corrections need special piecemeal justifications. And of course, not even the most garrulous self-obsessed bore is disposed to avow all their psychological history, but, plausibly, we are disposed, in normal circumstances, to avow any element of our psychological history.

For Tom to pass the criterion for psychological realism, as now understood, we need to show that, if we construct a putative psychological history or partial history of Tom, and Tom is not disposed to avow it, then that is sufficient warrant to reject the proffered history. Provided, of course, that Tom is also prepared to accept some alternative psychological explanation of his behaviour.

The contrast between Tom and our dumb friends is obvious: Tom cannot fail the criterion for psychological realism, provided he is disposed to accept some psychological explanation of his behaviour, but our dumb
friends not only cannot pass it but they cannot fail it either - in their case we have no test for when, and when not, two putative histories each sufficient to explain the creature's behaviour but imputing different overall conceptual contents are parts of the same overall psychological history.

The remaining business of this section is to show why, if Tom is not disposed to accept a proffered psychological explanation, then that is sufficient warrant to reject that putative explanation. That is to say, the remaining business of this section is to show why we have, to the extent that we do have, privileged access to our own volitions and cognitions.

Davidson developed his paratactic analysis for sentences of the form

Galileo said that the earth moves.

but he intended that analysis as a prototype for reports of propositional attitudes in general. The immediate problem which such sentences presented for Davidson was how can their truth conditions be determined by a finitely specifiable truth theory, since they are clearly infinite in number and yet "Galileo said that..." is equally clearly not a truth functional connective? Now the paratactic analysis solved this problem because it saw the utterance as two distinct utterances, of which the first is the noncompound sentence "Galileo said that" to which the truth theory assigned the primitive relation of samesaying, a relation with an infinite domain and range.
In this respect it is no more or less problematic for a finitely specifiable truth theory than other primitive relational terms; say, "longer than", which also has an infinite range and domain.

But Davidson's philosophical aim was wider than merely accommodating utterances of the form

Galileo said that the earth moved.

in a finitely specifiable truth theory. He intended thereby to explain how we creatures with finite understandings could grasp the meanings of any of an indefinite number of such sentences. Thus he wrote that a truth-theoretical account of logical form must:

\[\text{demonstrate in a persuasive way that the language, though it consists in an indefinitely large number of sentences, can be comprehended by a creature with finite powers.}\]

In this project his paratactic analysis is only a first step. The domain and range of the primitive term "samesays" are infinite and so cannot be comprehended in extension by creatures with finite powers. We need to provide a finitely graspable sense or intention which, in context, determines this infinite extension.

Compare, for example, the primitive relational term "longer than", also infinite in range and domain. Clearly, we finite creatures cannot grasp its infinite range and domain in extension, but we understand the term by grasping its intention or sense, an intention which, in context, determines its infinite range and domain. And it seems clear that one grasps the sense of "longer than" by learning, in the first instance, an operational test for
determining of arbitrary distances A and B whether or not
A is longer than B, or perhaps some finite set of such
operational tests - e.g. laying-off a ruler and counting,
taking longer to walk A than to walk B, and the like.
Now, plausibly, one only grasps the sense of "longer than"
when one grasps, along with the operational tests, a theory
which relates length to other properties, which explains
why these various operational tests determine the same
relation. And clearly, as the theory grows we apply the
relation to cases to which the original operational tests
cannot be applied: the distance from the nucleus to the
outer shell is greater than the distance from the nucleus
to the inner shell of electrons. And finally, given such
a developed concept, I do not want to claim a clear, sharp
or determinate boundary between what is constitutive of
the sense of "longer than" and what are contingent facts
about the relation. But without a finite theory involving
the relation, a theory which is grounded in some operational
tests, we could not grasp the sense of "longer than",
whether or not that sense can be determinately separated
from the contingent facts recorded in the theory.

Returning to samesaying, we do have a graspable
operational test. If by uttering "Galileo said that the
earth moves" I make Galileo and myself samesayers, it will
be because Galileo uttered some token with the same
meaning as the second of my utterances - where, for
Davidson, our respective tokens have the same meaning if
and only if they have the same truth conditions, more or less.
However, when we turn to belief, the trail blazed by Davidson grows cold. We need some finitely graspable way of identifying a belief state's functional role in order to select the appropriate index for it. And here we encounter two problems: (i) a functional role is holistically defined by its inter-relations with all other functional states, and (ii) we cannot survey the interplay of Tom's relevant functional states, but have to reconstruct fragments of them from his observed behaviour and the environmental context of that behaviour. Since we cannot survey the holistically inter-related functional system we need some operational test, some way of breaking into this hermeneutic circle, if we are to ascribe propositional attitudes.

A graspable way-in is provided if a belief-assertion generalisation, something of the form

A normal competent English speaker who is not reticent believes that \( p \) iiff he is disposed to sincerely reflectively assert that \( p \),

has a priori status. We can attribute propositional attitudes only if we are self-attributers with privileged access to our own minds. And we have a real cognitive and volitional life only because we are such creatures. True enough, Tom crouching in the cupboard under the stairs will not move a muscle, far less assent to the proposition that you are an escaped psychopath, but Tom's circumstances are exceptional. And Dick will never give voice to his secret fears of private anatomical shortcomings, but he is reticent on the matter. In such cases we ascribe belief on collateral evidence alone, and despite the absence of
reflective assent. But collateral evidence is unsystematic and unsurveyable in any way that could explain our grasp of the believing relation. Just so we always ascribe beliefs to dumb animals on collateral evidence, and have nothing to fix the ascriptions with any conceptual precision.

Furthermore, the institution of language itself gives an a priori guarantee of the belief-assertion generalisation. The general point of asserting is to change the belief state of the audience, and, in general, an audience so changes his belief state because he takes the utterer to believe what he says. The communicative point of uttering p is normally acheived because the audience thereby takes an utterer to believe that p, and takes the utterer to be justified in that belief. So, a language which includes reports of the beliefs of the language users presupposes an assertion-belief generalisation, and presupposes that its truth is general knowlege.

But so far, it would seem, any assertion-belief generalisation would do, so long as it was known to hold by the language users. Suppose, for example, F(p) was a computible and humanly graspable function taking the conceptual content of p onto some other conceptual content (for example, the negation function), then the following would serve communication intentions:-

A normal English speaker who is not reticent will be disposed to sincerely reflectively assert that p iff he believes that Fp.

Communication would be acheived because an audience, on
hearing an utterance p would compute the conceptual content F(p), and hence adjust his belief-state accordingly. And the utterer, knowing all this, would choose his utterance accordingly. But, it seems, since F( ) is graspable and defined on the conceptual contents of the primitive terms and syntactical operators of the language, this whole process is tantamount to the community assigning to the token p the conceptual content F(p) in the first place, and using the 'straight' assertion-belief generalisation. Hence the status of that generalisation is a priori.

Could a Martian, viewing with X-ray eyes the passing show of neural events within human skulls, index the same inter-connected set of second-order events centred on Tom's skull by a different set of conceptual contents, different from those by which Tom and his fellows index them? If so, the mere fact that Tom avows one set of conceptual contents but not the other would not be sufficient to show that those states really had the one set of conceptual contents and not the other. But, given that the sentences by which Tom does index his internal states do have determinate conceptual contents, I do not see how there could even be the possibility of such a re-indexing. My ground for this last claim is naturalism: physical facts determine conceptual facts. I do not see how the sentences Tom and his peers utter could have the one system of conceptual contents, if the other system of conceptual contents equally indexed their states. Since
both conceptual systems, *ex hypothesi*, equally index the same second-order physical facts, there would be nothing about those second-order physical facts to determine the one system of conceptual facts rather than the other, and their utterances could not determinately manifest the one set of conceptual contents rather than the other. Nor could there be a gap between a *plurality* of conceptual systems which equally index the *individual* Tom's functional states, and the *single* system of conceptual contents manifested by public usage of their language. For I am supposing that Tom *grasps* the public meanings, the determinate public concepts, and there is nothing over and above his individual inter-related second-order physical states to realise his grasp of those concepts.

What does it take to be a competent English speaker? How detailed does Tom's grasp of the relevant concepts need to be for his assertions to imply the corresponding beliefs? I am inclined to be generous. If Tom says that he believes in Einstein's general relativity, but his explanation is inadequate or confused, we *could* see public language as fragmenting into an overlapping set of dialects, and Tom's utterances manifesting non-standard beliefs. But I am inclined to follow Putnam's remarks about the division of linguistic labour\(^\text{21}\) in the other direction. Tom intends his utterances about relativity to have the same truth conditions as those of the cognoscenti. To attribute to Tom now the concept of relativity is to see the relevant functional system as embracing how he would respond to
instruction by the cognoscenti - he only has to pass the exam, not get 100%. Since the functional physical facts which determine conceptual content will include facts about how he would respond if ..., I can see no reason for a blanket exclusion of how he would respond to, say, reading a book on physics. Of course, he may prove incapable of improvement, or respond by saying that if that's Einstein's theory he doesn't believe it. Then we would have no warrant to infer from his earlier assertions that he did believe in relativity.

Section (vi): An assertion-belief generalisation

In this section I will attempt to formulate an assertion-belief generalisation which provides an operational test and for ascribing beliefs/which is correct in general but defeasible in any particular application - i.e. any single application of in may lead to an incorrect ascription of belief, but most applications of it must produce correct ascriptions of belief. I take it as no objection to what follows that my assertion-belief generalisation (A-BG) - unlike, e.g., Kripke's disquotation schemas, see Appendix B to this chapter - is not formulated in Tom's home language. There is no reason why an operational test for asserting sentences of the form X believes that p should be expressible in that language.

Kripke considers the following disquotation schema$^{22}$:-

A normal English speaker who is not reticent will be disposed to sincere reflective assent to "p" iff he believes that p.
Kripke finds the disquotation schema problematic for reasons which I discuss and reject in Appendix B, but there is another reason why we cannot adopt his disquotation schema: the contexts of the believer and the belief-reporter may be significantly different. For example, Suppose I report to you that Tom believes that I am his father. Tom, looking me in the eye, sincerely and reflectively produced a token of "You are my father", and nodded gravely when I produced a token of "I am your father". But this is not correctly reported by me, to you, as "Tom believes that you are my father" nor as "Tom believes that I am your father"!

Indexicals, in the token to which Tom assents and in the corresponding token in the belief report, differ because of the different contexts of believer and belief-reporter.

Perhaps we can avoid this problem by taking Tom to be sincerely disposed to give reflective assent to the token employed in the belief report, i.e., to the second half of Tom believes that I am his father?

Unfortunately not. Suppose, sitting in Tom's dingy bedsitter in London, I correctly report to him

Your mother believes you are doing well in Australia. She is not disposed to give sincere reflective assent to the second half of my utterance. If she were aware of that token, she would not assent to it, given the circumstances of its utterance. Perhaps we could try: If she were aware of the token but not of the circumstances of its utterance, she would assent to it? But she has to be aware of the context to know the values of the indexicals employed, that it is whey-faced Tom who is being addressed. There
seems no practical way of allowing her access to the values of the indexicals in the belief report whilst screening off all features of the context that might cause a 'change of mind'. And since we are seeking an operational test for ascribing beliefs, merely theoretical possibilities are irrelevant.

Consider, instead, the following sketch of a generalisation linking attributions of belief to assertions of sentence tokens:

\[(x)(p)[Nx \rightarrow (x \text{ believes } p \leftrightarrow (\exists t)(Axt \& s(t)=s(p) \& ...))\]  

Here "(x)" quantifies over psychological subjects, "(p)" and "(\exists t)" over token utterances of English. "Nx" means that x is a normal competent English speaker. "x believes p" contains the variable p, since we are now presupposing a paratactic analysis of, e.g., "Tom believes that the earth is round" and "p" replaces the demonstrative "that". "Axt" means that x is disposed to give sincere reflective assent to t. And "s(p)=s(t)" means that t and p have the same semantic content. We need to insert a further condition linking t and p in the gap ..., for otherwise attributions of belief will become transparent, requiring only that content clauses have the same semantic content.

We want to permit, as legitimate transitions from t's to p's, transitions from "You are my father" to "I am his father", and from "It will rain" to "It did rain", in suitable contexts, but not from "Cicero denounced Cataline" to "Tully denounced Cataline". We need to screen off those singular terms, like proper names, which may not, in
general, change in going from a t to a p, from those singular terms, like indexicals, which may. But even indexicals are not fully transparent. If, for example, Tom assents to the token

He bit her and he scratched her.

and we happen to know that the two occurrences of "he" have the same value, we may be wrong to report

Tom believes he bit and scratched her.
because Tom might believe the biter and the scratcher to be two different persons.

I suggest that we append numerical subscripts to the singular terms (including, e.g. tenses) in t which may be changed in a belief-report p, and the corresponding singular terms in p. Where two such singular terms may be taken by a competent, reflective asserter of t to have different semantic values they are to be given different subscripts. And p likewise. (Thus "He hit him" would yield "He₁ hit him₂", but "He hit himself" yields "He₁ hit himself₁") And indexicals in p and in t are to have the same subscript only if they have the same semantic value.

If s is such a subscripting of the indexical terms in t and in p, let Pps and Pts denote the resulting subscripted sentence types. We can treat indexicals bearing the same subscript as occurrences of the same singular term and require Pps ⊨ Pts. I.e., the condition needed to complete the assertion-belief generalisation above is: There is a subscripting of the indexicals in t and p, as described above, such that Pps ⊨ Pts. My A-B G
is then:

$$(x)(p) \{ N_x \rightarrow (x \text{ believes } p \leftrightarrow (\exists t)(Axt \land s(p) = s(t) \land (\exists s)(Pps \vdash Pts))) \}$$

As an operational test, this A-B G should be pragmatically restricted to easily comprehensible English sentences. We can well imagine an English sentence which is generally assented to on the authority of some sage, but whose sense is too complex for even the majority of competent English speakers to grasp. And we probably have systems of belief so complex that, were they to be articulated in English, we would fail to assent to the resulting sentence. (The grammar of English?)

The generalisation is specific to English not for the trivial reason that I specified the values of "p" and "t" to be tokens of English, nor because "Nx" says nothing about competence in other tongues. Rather, the generalisation is specific to English because the syntactic rules which determine whether or not Pps $\vdash$ Pts are language specific.
Appendix A: Quine, Davidson and psychological realism

My view that if a set of physical facts is to determine a set of psychological facts, then it must determine one set of psychological facts, should be congenial to Quine and perhaps also to Davidson. Quine famously argues from the indeterminacy of radical translation, where translation is constrained by the totality of natural (behaviouristic) facts, to the radical indeterminacy of interpretation of a brother speaker of the home language. And this motivates his "flight from intention" and his explication of propositional attitudes as relations to eternal sentences. It is best, I think, to view these eternal sentences as indices of the subject's natural states, in the manner of Loar. So understood, Quine is led by the indeterminacy of the conceptual content of a psychological history to jettison psychological facts from his ontology and to explicate propositional attitude reports as sentential indices of the physical (behavioural) facts. Davidson's position is less clear to me. He seems to find a double indeterminacy: firstly, different assignments of concepts and extensions to the primitive vocabulary of a language yield the same truth conditions for complete sentences, and secondly, assigning truth conditions to the native utterances is not independent of assigning beliefs and other propositional attitudes to native utterers, where significantly different assignments may be holistically consistent with the behavioural evidence.
Davidson's position is unclear to me because he appears to regard these two outbreaks of indeterminacy as having different ontological significance: the indeterminacy of sub-sentential semantics relative to sentential semantics leads Davidson to accord a merely instrumental status to assignments of meanings and extensions to the primitive vocabulary of a language, but the indeterminacy of assignments of truth conditions to beliefs and utterances relative to the behavioural whole does not seem to warrant, in Davidson's view, a parallel ontological downgrading of the assignments of truth conditions to beliefs and to utterances. This difference may be because Davidson thinks the two indeterminacies are of different kinds: the indeterminacy of sub-sentential relative to sentential semantics is an indeterminacy in principle, given one complete truth theory for an extensional language we could easily manufacture a second theory assigning different extensions to the primitive vocabulary but yielding the same crop of truth conditions for whole sentences. On the other hand, the indeterminacy of assignments of truth conditions to the native beliefs and utterances relative to native behaviour may merely be a case of the ubiquitous underdetermination of theory by finite empirical evidence having no special ontological significance.

However, my criterion for psychological realism – one psychological history to a subject – clearly presupposes a rejection of Quine's thesis of the indeterminacy of radical translation. For if there were such Quinean
indeterminacy, then, given a psychological history (set of sentences) of Tom, we could translate it, first using one translation manual and then another, to produce two different psychological histories of Tom between which the physical facts were powerless to adjudicate, thus destroying my criterion.

I am saving my main discussion of Quine's and others' arguments for the indeterminacy of radical translation for chapter three, but I propose to reject here three ways of attempting to counter Quine's arguments.

Firstly, it is not open to me to retort that the two psychological stories would be over-all conceptually equivalent, and therefore both reports of the same psychological facts. For the strategic intent of Quine's argument was to banish such psychological facts from our ontology, to confine reformed propositional-attitude reports (reporting attitudes to eternal sentences) to the role of merely indexing physical states and their interrelations. So the alternative translations envisaged by Quine will have intuitively different conceptual contents: what it is like to think that this rabbit is eating lettuce is intuitively different from what it is like to think that this locally bounded proper part of the concrete universal rabbithood is eating lettuce, which is intuitively different from ... The philosophical interest of indeterminacy arguments lies in their power to deprive us of the belief that there is a fact of the matter as to what it is like, cognitive and volitionally, to think Tom's
thoughts. Quinean indeterminacy would achieve this end by showing that the physical facts always support at least two psychological histories (sets of sentences) which, intuitively, assign different conceptual contents to Tom's thoughts.

Secondly, it would not be a reply to Quine to concede that more than one system of sentential indices fit the physical facts, ascribing different conceptual contents, but claim that we simply choose, by a logically arbitrary convention, one such system of indices, and incorporate our choice into the very meanings of our psychological verbs "believes", "desires", etc., so that it becomes uniquely correct to say of young Tom tending his rabbit "He believes his rabbit is eating lettuce", and not to say of him "He believes that this locally bounded proper part of the concrete universal rabbithood is eating lettuce", although it would be equally correct to say, in a different sense of "believes", "Tom believes that this locally bounded proper part etc.". And finally to claim, pace Quine, that the same set of physical facts can determine simultaneously both these non-competing psychological histories. Tom, uncorrupted by philosophical studies, simply does not realise that he also believes that this locally bounded proper part etc., because he is adept at reporting his own and others' psychological states using the concepts of belief, desire, etc. rather than the concepts of belief, desire, etc., and is thereby party to the convention of using "rabbit"-sentences
as indices rather than "rabbithood"-sentences as indices.

This reply will not do because it illegitimately tries to isolate the translation of a native's psychological verbs from the translation of the sentences they govern. No doubt one radical translator might translate a given native sentence as

Tom believes that this rabbit is eating lettuce and hold that the natives have adopted the convention of so indexing that internal state and built that convention into the very meanings of their psychological verbs. But, if Quine's or others' arguments for the indeterminacy of radical translation succeed, a second radical translator could translate that same native sentence as

Tom believes that this locally bounded proper part etc. and hold that the natives have conventionally adopted a different indexing system, and built that convention into the very meanings of their psychological verbs. If this is so for the translation of native sentences, then it is also so for the 'domestic' case of interpreting the propositional attitude reports of a brother speaker of the home language.

Supposing that translation between (interpretation of) the utterances of speakers of the home language is indeterminate, there can be a convention among speakers of Home that Tom's attitude is correctly reported by the sentence

Tom believes the rabbit is eating lettuce but not a convention that this sentence expresses in Home the conceptual content that Tom believes that the rabbit
is eating lettuce, rather than the conceptual content that Tom believes\(^1\) that this locally bounded proper part etc.
The point is, following Lewis\(^2\), that a convention among a population is a regularity of action (a correlation between their actions and something else) which is mutually known to obtain by the members of the community, and which continues to obtain because they expect it to. The Home language speakers can mutually know that Tom’s feedings of his pet are correlated with a certain sentence in Home, but not, if Quine is right, that the sentence is correlated with a certain conceptual content. So they cannot choose one conceptual content rather than another by convention.

Thirdly, some may feel that my broadly Quinean naturalism, and the derived criterion for psychological realism, are too strong. Faced with Quinean indeterminacy of the conceptual contents of utterances relative to the physical facts of the matter, and the related indeterminacy of the psychological history of Tom relative to his physical history, some may feel that Quine’s flight from intention was over reacting. Their naturalism requires only that the conceptual content of utterances, the psychological facts about Tom, are determined by the relevant physical facts plus an interpretation of those facts, and Quinean indeterminacy merely reflects that the physical facts may be now interpreted one way and then reinterpreted in some other way. But such a view would be naturalistic only in name. A non-naturalist may agree, for example, that
action in the physical world requires the agent to perform a physical movement, but what makes the action a non-natural event is that it also requires some additional non-natural factor, and that factor is the intentional interpretation. Furthermore, my cognitive and volitional life does not seem to occur uninterpreted: when not befuddled by philosophising I know the difference between contemplating a rabbit eating lettuce and contemplating a rabbit feature destroying a lettuce feature. And the fact that the one 'interpretation' rather than the other is incorporated into my cognitive and volitional life itself requires to be grounded in the physical facts about me, in any naturalism worthy of the name. So I return to my Quinean naturalism, my rejection of Quinean indeterminacy, and my criterion of psychological realism.
Appendix B: Kripke's doubts about his disquotation schema

Kripke is inclined to accord self-evident a priori status to his disquotation schema: -

A normal English speaker who is not reticent will be disposed to sincere reflective assent to "p" iff he believes that p where substitutions on "p" are confined to sentences of English. Kripke hesitates to endorse the schema because, given a case he constructs, it leads to a contradiction. We shall see that applying my A-B G to Kripke's case also generates a contradiction. Resolving that contradiction reveals another problem, which Kripke also identifies. Solving this second problem by distinguishing semantic and syntactic inconsistency begins to show how propositional attitude reports fit together in psychological explanations.

But first Kripke's example. Briefly, Pierre, a normal monoglot French speaker who has never visited London comes to sincerely reflectively assent to the sentence "Londres est jolie", perhaps as a result of travellers' tales, holiday brochures and the like. The corresponding disquotation schema for French sentences, and normal principles of translation from French to English, would lead us to assert that Pierre believed that London was pretty. Pierre then moves to London, but does not realise that it is the city he in fact refers to as "Londres", perhaps because he is now incarcerated in some unsalubrious district, does not realise he is now living in the capital of England, the city containing Buckingham Palace, or whatever. Pierre learns English by the direct method,
learns that the city he is now in is called "London", sincerely, reflectively assents to

Londres est jolie

but not to

London is pretty.

So, by the disquotation schemas for French and for English, and the normal rules of translation, we are committed to asserting both

Pierre believes that London is pretty
Pierre does not believe that London is pretty

and thus we contradict ourselves. Hence Kripke's hesitation over endorsing his otherwise plausible disquotation schema.

This contradiction would be harmless if Pierre's total behaviour could be sub-divided into two mutually exclusive subsets, one of which involved the belief that London is pretty and the other not, the other might even involve the belief that London is not pretty. For surely we should not expect functional theory, when applied to a given subject, to integrate all his attitudes into one consistent functionally inter-related set. Rather, we should expect functional theory to integrate any single attitude into an inter-related set of attitudes, but the resulting sets may be partially overlapping or even mutually disjoint. And there need be no functional explanation why I hold one set of attitudes in one context and another set of attitudes, inconsistent with the first, in another context.

An analogy: the functional theory of the motor car may explain my car's behaviour from $t_1$ to $t_2$, and from $t_3$ to $t_4$. 
but not what was going on between \( t_2 \) and \( t_3 \), when it was being stripped down and re-assembled in the garage. Perhaps Tom, too, undergoes a radical change of belief and personality caused by a light tap with a hammer on the base of his skull, and there is no psychological theory spanning the discontinuity and explaining the changes by integrating them into one rational pattern.

What now of my criterion of psychological realism, that there be one psychological history of Tom? The present point does not affect that criterion. What it is like to be Tom, the true psychological history of Tom, can be peppered with psychological discontinuities, irrational changes of belief and mood from context to context, without there being a second psychological history of Tom assigning a different set of conceptual contents to his attitudes. The application of psychological theory to Tom must yield a determinate set of attitudes and conceptual contents, be they as quirkily discontinuous as they may.

However, we can easily block this escape route in Pierre's case. Come the holidays, we find Pierre entering the forecourt of a London station, suitcase in hand. The intention of his action is to leave London for Londres - if I may so put it. This intention is motivated by his not believing that London is pretty (so he is quitting London) and by his believing that Londres is pretty (so he is going to Londres). His bodily movement is not two actions - leaving London and going to Londres - with disjoint motivations, like the same bodily motion might
be both waving to a friend and hailing a taxi, where each motive is separate and sufficient to explain the behaviour. Neither the intention to leave London nor the intention to go to Londres is individually sufficient to explain Pierre's bodily motion. Both are needed to explain his heading out of London in that direction. Hence our problem is to give a consistent explanation of Pierre's attitudes in relation to this one action.

Pierre's case generates a contradiction from my A-BG also. Pierre is not disposed to assent to tokens of "London is pretty" whence the A-BG licenses us to draw the conclusion that he does not believe London is pretty. But an A-BG for French and his disposition to assent to tokens of "Londres est jolie", plus the usual rules of translation from French to English permit us to draw the conclusion that he does believe that London is pretty, and we have contradicted ourselves.

The way out of this contradiction is, I think, to argue that "London is pretty" and "Londres est jolie" are not mutual translations: in fact, proper names in one language are untranslatable into another language, that is what Pierre's predicament shows. Translation is required to preserve conceptual content, and Frege's criterion of conceptual content was that X and Y have the same conceptual content iff one who grasps X and grasps Y cannot believe X and not believe Y, or vice versa. On this philosophical criterion of conceptual content Pierre's case shows precisely that, just as no two names within a language
have the same conceptual content, so no two names in
different languages have the same conceptual content, no
matter that the etymology of "Londres" traces it back to
"London", and no matter that the non-philosophical
conception of translation requires every French word to
disappear in favour of English words, and hence "Londres"
to be replaced by "London".

On the Fregean, philosophical view of translation
Pierre does not believe that London is pretty, indeed he
believes that London is not pretty. But he also believes
that Londres is pretty. Now, it seems, we do not contradict
ourselves when attributing attitudes to Pierre, rather we
attribute inconsistent beliefs to Pierre. The inconsistency
lies in Pierre's beliefs, not in our attributions of belief
to him.

However, Kripke cannot accept that Pierre has
inconsistent beliefs. He writes:-

Prima facie, we should distinguish semantic inconsistency
(P,Q |-- *) from syntactic inconsistency (P,Q |-- *). Pierre's
beliefs are syntactically, and therefore semantically,
inconsistent if and only if he would express them by
syntactically inconsistent utterances. But they are only semantically inconsistent if he would express them by syntactically consistent sentences. Hopefully Pierre is the victim of semantic but not of syntactic inconsistency, so understood. And to have semantically but not syntactically inconsistent beliefs casts no slur on Pierre's rationality - there is an epistemically possible world in which all his beliefs are true together. Further, only syntactic inconsistency is discoverable a priori, not the more insidious semantic inconsistency of beliefs. 29

Nonetheless, attributing semantically inconsistent beliefs to Pierre will not, by itself, carry us clear of the wood. Pierre assents to tokens of "That is a pretty city" and to tokens of "This is not a pretty city", where the indexicals both refer to London. The A-BG, which permits changes of indexicals to suit the changing contexts of believer and belief-reporter, will yield the belief reports:

Pierre believes of London that it is a pretty city
Pierre believes of London that it is not a pretty city.

Now Pierre's beliefs are reported by clauses which are syntactically inconsistent one with the other. We attribute beliefs to Pierre in order that, together with the desires we also attribute to him, we thereby explain his intentions. And what relates the sentences we employ to index his beliefs and desires to one another, and to the sentences by which we index his intentions, are inference rules. The problem is: If we index his semantically inconsistent
beliefs by syntactically inconsistent indices, then all and any intention will be rational, because syntactically inconsistent sentences imply every sentence. So, it seems, our attributing syntactically inconsistent indices to Pierre's beliefs can play no role (or a vacuous role) in our explanations of Pierre's actions. But, finally, if syntactically inconsistent indices play no role in explaining Pierre's actions, we can have no warrant to apply such indices to Pierre's beliefs in the first place!

Let me emphasise: We cannot avoid the problem merely by confining ourselves to mutually consistent indices. The A-BG warrants indexing his beliefs by p and also by -p, those indices fed into the inferential machinery will generate all others. So whilst we might show, by confining ourselves to mutually syntactically consistent indices, that Pierre walking into the station forecourt is rationally motivated, that is only because all patterns of motivation, rational and lunatic, motivate Pierre.

Can we solve the problem by making \( \vdash \) a relevance logic, thereby ensuring that contradictory premisses do not imply every sentence? This would help, but not much: given syntactically inconsistent indices as premisses, the rules of inference, even of a relevance logic, are still going to have a hayday churning out syntactically inconsistent indices for our subject's beliefs - for example, that Pierre believes of London that it is pretty and not pretty. And Pierre is a master logician; his beliefs and other
attitudes seem to him to be individually and collectively perfectly consistent, he being the very model of a rational mind. But the indices we ascribe would not show this, they would present his beliefs as syntactically inconsistent, his mind as irrational.

The solution lies, I think, not in trying to modify the A-BG, but in imposing rules on how we link up and mutually orient individual attitude reports into action expaining complexes. To proceed I need to anticipate the discussion to come in chapter five. We attribute perceptions, beliefs and desires in order to explain intentions, and in chapter five I shall argue that to link up perceptions, belief, etc. reports into an explanatory complex, and to orient one index relative to the others, we shall need some version of quantification across psychological verbs. But we need to prevent that quantification from inducing transparency. We can achieve a mode of quantification-in which performs the tasks of linking up and mutually orienting the various indices so as to form an explanatory complex, but without inducing unwanted transparency, by developing the paratactic analysis of propositional attitude reports to yield an associated conception of quasi-binding the indexical elements in the indices.

From the individual attitude reports which are to be united into an explanatory complex we need to fashion maximal propositional functions. The idea is most easily grasped from an example. From

Tom desires that Dick hit Harry and Tom believes that Dick hit himself
we form

\[
\begin{align*}
\text{Tom desires } & [\text{Dick } x_1 & \& x_1 \text{ hit } x_2 & \& \text{Harry } x_2] \\
\text{Tom believes } & [\text{Dick } x_3 & \& x_3 \text{ hit } x_3]
\end{align*}
\]

where "Dick \(x_1\)" etc. are new-minted predicates, and no variable in one clause is repeated in another. Then to link up and orient the indices we quantify across and add identity beliefs as, for example

\[
(\exists x_1, x_2, x_3)(\text{Tom desires } [\text{Dick } x_1 & \& x_1 \text{ hit } x_2 & \& \text{Harry } x_2] \\
& \& \text{Tom believes } [\text{Dick } x_3 & \& x_3 \text{ hit } x_3] \\
& \& \text{Tom believes } [x_1 \neq x_2 & \& x_1 = x_3])
\]

The variables peculiar to each of the maximal propositional functions are to be repeated only in the identity index. We hope then to take the various indices to be related to one another in a formal, syntactic, context-free manner as grist to the mill of inference rules which accept open as well as closed sentences.

Of course, this project is threatened by transparency. We have

\[
(\exists x_1, x_2)(x_1 = x_2 & \& \text{Pierre believes}[P x_1 & \& -P x_2])
\]

How can we prevent this implying, of master logician Pierre,

\[
(\exists x)(\text{Pierre believes}[P x & \& -P x])?
\]

The answer lies in a paratactic analysis of the indices and the quasi-binding of the variables therein.

Consider first

\[
\text{Tom believes of } \text{Jack and Jill that she hit him}
\]

and let us rewrite this in the style

\[
\text{Tom believes of } \langle a_1, b_2 \rangle \text{ that } [x_1 \text{ hit } x_2]
\]

where "[...]" is the paratactic clause and the occurrences of "a" and "b" in "\(\langle\ldots\rangle\)" are transparent. The paratactic
clause is an interpreted sentence, and the subscripts in "[...]" are demonstratives whose values are the values of the correspondingly subscripted members of "\langle...\rangle". The values of "x_1" and "x_2" in "[...]" are thus a and b, respectively. We can quantify into "\langle...\rangle" to get say

\((\exists y,z)(\text{Tom believes of} \langle y_1,z_2 \rangle \text{ that} [x_1 \text{ hit } x_2])\)

where I have used new variables to show that the quantifiers do not bind the variables "x_1" and "x_2" in "[...]". Now the paratactic clause is as open sentence: the values of "x_1" and "x_2" are shown to be the first and second members of "\langle...\rangle", but "\langle...\rangle" contains only bound variables, thus no values have been assigned to "x_1" and "x_2" in "[...]". A specific belief state is indexed by an open sentence only relative to an assignment of semantic values to its free variables. But then, of course, the index is not an open sentential function but a closed sentence. So the sentence above, being open, does not index a specific belief state of Tom's. Rather, it says, in effect, that Tom is in a belief state of a certain type.

Now consider

\((\exists y,z)(y=z \& \text{Tom believes of} \langle y_1,z_2 \rangle \text{ that} [Fx_1x_2])\)

This implies

\((\exists y)(\text{Tom believes of} \langle y_1,y_2 \rangle \text{ that} [Fx_1x_2])\)

but not the perhaps irrational

\((\exists y)(\text{Tom believes of} \langle y_1 \rangle \text{ that} [Fx_1x_1])\)

And clearly, also, if we have

\((\exists y)(y=a \& \text{Tom believes of} \langle y_1 \rangle \text{ that} [Fx_1])\)

we may infer
Tom believes of \( a_1 \) that \([ Fx_1]\)

but NOT

Tom believes that \([ Fa]\)

The forbidden inferences would move from an index of the form \( Fx_1x_2 \) where "\( x_1 \)" and "\( x_2 \)" have the same semantic value, to an index of the form \( Fx_1x_1 \), and moves from an index of the form \( Fx \) where the semantic value of "\( x \)" is a to an index of the form \( Fa \).

Although the paratactic analysis has given us the means to block the generation of syntactically inconsistent indices, we have not yet seen how the external quantifiers can link up and mutually orient the various indices in an explanatory complex. So far, the external quantifiers do not even quasi-bind the variables in indices.

Let us return to Tom, Dick and Harry. Suppose the following is the case:-

\[(\exists yzw)(y=z=w \& Tom desires of \langle y_1, z_2 \rangle that [Dickx_1 \& x_1 \text{hit} x_2 \& Harryx_2])
\]
\[\& Tom believes of \langle x_3 \rangle that [Dickx_3 \& x_3 \text{hit} x_3]
\]
\[\& Tom believes of \langle y_1, z_2, w_3 \rangle that [x_1 \neq x_2 \& x_1 = x_3]
\]

We could let the subscripts in the praratactic clauses show Tom's beliefs about the indentities and differences of the semantic values of his representational states so reported, and then feed the clauses, complete with subscripted variables, into the mill of context free, syntactically defined inference rules. Whence, since

\[Dickx_1 \& x_1 \text{hit} x_2 \& Harryx_2, x_1 \neq x_2 \& x_1 = x_3 \vdash Dickx_3 \& x_3 \text{hit} x_3\]

is not a valid sequent, we show that Tom has not yet attained his heart's desire.
But the way in which I described what it is that the subscripts show about Tom's attitudes was misleading: it made it appear that they show a higher-order belief of Tom's, a belief about his other attitudes. Whereas, I shall argue in chapter five, Tom's beliefs about identities and differences, avowed by him and part of a history which can be realistically interpreted, are not beliefs about beliefs etc., but beliefs of the same order and about the same entities as those other beliefs etc.

We can show this by employing the same (syntactically) variables in the quantifiers and in the "⟨..⟩" clauses as in the "[...]" clauses, as follows:

\((\exists x_1 x_2 x_3) (x_1 = x_2 = x_3 \& \text{Tom desires of } \langle x_1, x_2 \rangle \text{ that } [\text{Dick} x_1 \& x_1 \text{hit} x_2 \& \text{Harry} x_2])\)

\& Tom believes of \(x_3\) that \([\text{Dick} x_3 \& x_3 \text{hit} x_3]\)

\& Tom believes of \(\langle x_1, x_2, x_3 \rangle\) that \([x_1 \neq x_2 \& x_1 = x_3]\)

This, roughly, is what I mean by the external quantifiers quasi-binding the variables in the paratactic clauses. I hope to show more precisely is meant by developing a truth theory for such quantification in chapter five Section (iv).

I take it that, the variables in "⟨..⟩" being transparent, the above implies

\((\exists x) (\text{Tom desires of } \langle x_1, x_2 \rangle \text{ that } [\text{Dick} x_1 \& x_1 \text{hit} x_2 \& \text{Harry} x_2])\)

\& Tom believes of \(x_3\) that \([\text{Dick} x_3 \& x_3 \text{hit} x_3]\)

\& Tom believes of \(\langle x_1, x_2, x_3 \rangle\) that \([x_1 \neq x_2 \& x_1 = x_3]\)

where all the "\(x_1\)" in all the "⟨..⟩" are recurrences of the same bound variable. But the latter does not imply the former, because the subscripts in one paratactic clause of
of the latter bear no significant relation to the subscripts in its other paratactic clauses.

We now have a sufficient apparatus to record the semantic inconsistency of Pierre's beliefs, and also the rationality of his intention in entering the station forecourt. We have:

\[(\exists x_1 x_2)(x_1=x_2 \& \text{Pierre believes of } <x_1, x_2> \text{that } [P_{x_1} \& -P_{x_2} \& x_1 \neq x_2])\]

Call this (1). Further, Pierre's character is such that:

\[(x_1 x_2 x_3 x_4)(\text{Pierre believes of } <x_1, x_2, x_3, x_4> \text{that } [P_{x_1} \& -P_{x_2} \& x_1 \neq x_2 \& x_1 = x_3 \& x_2 = x_4])\]

\[\rightarrow \text{Pierre desires of } <x_3, x_4> \text{ that } [A_{x_3} \& -A_{x_4}]\]

Call this (2). We wish (1) and (2) to imply:

\[(\exists x_1 x_2 x_3 x_4)(\text{Pierre believes of } <x_1, x_2, x_3, x_4> \text{that } [P_{x_1} \& -P_{x_2} \& x_1 \neq x_2 \& x_1 = x_3 \& x_2 = x_4] \& \text{Pierre desires of } <x_3, x_4> \text{ that } [A_{x_3} \& -A_{x_4}])\]

Call this last (3). The ordinary rules of quantifier elimination will not permit different quasi-bound variables to be replaced by the same name. So those rules yield:

\[a=b \& \text{Pierre believes of } <a, b> \text{that } [P_a \& -P_b \& a \neq b] \] \[\rightarrow \text{Pierre desires of } <a, b> \text{ that } [A_{a} \& -A_{b}] \]

\[\text{Pierre believes of } <a, b, c, d> \text{ that } [P_{a} \& -P_{b} \& a \neq b \& a = c \& b = d] \]

\[\rightarrow \text{Pierre desires of } <c, d> \text{ that } [A_{c} \& -A_{d}] \]

Clearly the syntactic machinery requires the addition of a rule which allows us to go from (1) to:

\[(\exists x_1 x_2 x_3 x_4)(x_1=x_2 \& \text{Pierre believes of } <x_1, x_2, x_3, x_4> \text{that } [P_{x_1} \& -P_{x_2} \& x_1 \neq x_2 \& x_1 = x_3 \& x_2 = x_4])\]

Call this (1)'.

And equally clearly such a rule is plausible, since it is like going harmlessly from:

\[(\exists x_1 x_2)(\text{Pierre believes of } <x_1, x_2> \text{ that } [P_{x_1} x_2])\]
(\exists x_1 x_2) (\text{Pierre believes of } \langle x_1, x_2 \rangle \text{that } [Fx_1 x_2 \& x_1 = x_2 \& x_2 = x_2])

and thence to

(\exists x_1 x_2 x_3 x_4) (\text{Pierre believes of } \langle x_1, x_2, x_3, x_4 \rangle \text{that } [Fx_1 x_2 \& x_1 = x_3 \& x_2 = x_4])

So, our route to showing Pierre's rationality is from (1) to (1)'', then ordinary existential elimination on (1)'', followed by ordinary universal elimination on (2), followed by modus ponens and \&-introduction to yield:

\text{Pierre believes of } \langle a, b, c, d \rangle \text{that } [Pa \& Pb \& a \neq b \& a = c \& b = d]

& \text{Pierre desires of } \langle c, d \rangle \text{that } [Atc \& -Atd]

Then ordinary existential introduction yields (3).

Pierre being a master logician, we will want, in general, if \( P \models Q \), \( P, Q \) being open or closed sentences, then Pierre believes \([P] \Rightarrow \text{Pierre believes } [Q] \). And the identity rule \( a = b \), \( F(a) \models F(b) \) applies only where "a" is outwith the scope of a psychological verb. But of course it follows, where \( P \models Q \) depends upon this modified identity rule, we will have: Pierre believes \([P] \Rightarrow \text{Pierre believes } [Q] \).

I hope that the cat I hope to bell in chapter five has been let far enough out of the bag to give the prospect of a psychological theory which can (i) distinguish semantic inconsistency from syntactically consistent attitudes, so that those attitudes can function in rational explanations of behaviour, and which (ii) regiments the folk-explanations which, being acknowledged by their subjects, can be interpreted realistically.
Appendix C: McGinn on the relation between reference and belief

My A-BG was:-

\((x)(p)(Nx \rightarrow (x\text{believes} \Leftrightarrow (\exists t)(Axt \& s(p)=s(t) \& (\exists s)(P_{ps} \vdash P_{ts}))))\)

where "Nx" means that \(x\) is a normal competent English speaker, "Axt" means that \(x\) asserted token \(t\), "s(p)=s(t)" means that tokens \(p\) and \(t\) have the same semantic content, and "(\exists s)(P_{ps} \vdash P_{ts})" means that there is a subscripting \(s\) of the indexicals of \(p\) and \(t\) such that

(i) if someone can believe the values of two indexicals in \(p\) are different, then they receive different subscripts. And \(t\) likewise.

(ii) if an indexical in \(p\) and in \(t\) receive the same subscript, then they have the same value such that treating the subscripts as singular terms in place of the indexicals so subscripted, \(P_{ps}\) and \(P_{ts}\) are inter-derivable.

Are the belief reports, so generated, de dicto or de re? If total transparency is required for de re reports, then they are not de re. But if total opaqueness is required for reports de dicto, then neither are they de dicto. They are not totally transparent because, if Tom asserted that Cicero denounced Cataline, we are not thereby warranted to report that he believes Tully denounced Cataline. And if Tom asserts that he hit him, and we know that 'the he' was 'the him', so to speak, we are not thereby warranted to report that Tom believes that he hit himself. But, subject to this restriction, we are allowed to change indexicals to suit the change of context from...
believer to belief-reporter. Further, as we have seen in Appendix B and shall see more in Chapter Five, we need the A-BG to generate belief-reports that we can link up with other reports of attitudes to form explanatory complexes. This requires that a report of the form

\[ x \text{ believes that } F_{abc} \ldots \]

can be read as

\[ x \text{ believes of } \langle a_1, b_2, c_3 \ldots \rangle \text{ that } [F_{a_1 b_2 c_3} \ldots] \]

where the terms in "\langle \ldots \rangle" and in "[ ... ]" are open to quantification yielding

\[ (\exists x_1 x_2 x_3 \ldots) (\ldots x \text{ believes of } \langle x_1, x_2, x_3, \ldots \rangle \text{ that } [F_{x_1 x_2 x_3} \ldots]) \ldots \]

in which the variables in "[ ... ]" are bound and those in "\langle \ldots \rangle" quasi-bound.

McGinn produces a theory of reference which severs the link between, say, "Tom asserts "Cicero denounced Cataline"", and "Tom believes that Cicero denounced Cataline"; and between, say, "Tom asserts "You are my father"", and "Tom believes that I am his father." Deprived of this initial move from Tom's acts of assertion to reports of Tom's beliefs, we could not proceed to replace the singular terms in the belief-reports by bound and quasi-bound variables so as to form explanatory complexes. Or so it would seem. Thus McGinn's theory of reference, in effect, challenges my A-BG and the resulting psychological theory.

McGinn argues that reference is determined by linguistic rules plus the context of utterance: the referer need not have a thought about the referent. The plausibility of this claim is best seen from examples. A checker on a
car assembly-line, intent on the console of his monitoring apparatus may say

That car is certified as roadworthy
and thereby be referring to the particular car then occupying his station on the assembly line, but he need have no perceptual causal contact with the car, being intent on his instruments. Similarly, listening to a tale about someone he has never met and otherwise does not know, Tom might remark

He was very foolish
and thereby be referring to a particular person, determined by the context, but without having a singular thought about that person in particular.

The claim that singular reference is not sufficient for singular belief is tantamount to the claim that the relevant behaviour can be explained by general beliefs about whatever car was then ... or by general beliefs about whoever was the hero of the story. So although the subject does assert that the car is roadworthy (that he was very foolish), his behaviour is explained by the belief that whatever car is at his station is roadworthy (that whoever the story was about was very foolish). Thus the singular terms in the belief reports do not have the same semantic values as those in the assertions.

The beliefs which McGinn does ascribe to the utterers depends upon collateral information - that the utterer is a car assembly worker monitoring his console (that the utterer has been listening to a story). We wanted the
A-BG to provide an operational test leading from acts of assertion to reports of beliefs, to provide a route into the hermeneutic circle of psychological explanation, albeit a test which could, on any one occasion, be over-ruled and corrected by that theory. We wanted an a priori, context free, route from assertion to prima facie belief-report, not a route which depends upon knowing that in this case the asserter is an assembly-line worker, and in that case the auditor of a story. We return to such an a priori route if, when x asserts "Fabc...", we report

\[ x \text{ believes } [(\exists xyz...) (R''a''x & R''b''y & R''c''z & ... & Fxyz...)] \]

where "R''a''x" means that "a" refers to x. And further, since McGinn locates changing the belief-state of the audience as the general point of an act of assertion, the audience is, in general, warranted in believing that Fabc..., if he takes x to justifiably manifest the above belief by his act of assertion, and if the audience knows that R''a''a, R''b''b, etc.

An A-BG so modified, does yield a crop of res for quantification, binding and quasi-binding in explanatory complexes - viz., the referring expressions "a", "b", "c", ...

Moreover, these are public res, available to audience and asserter alike, if assertions are to serve their point. A narrow functionalist will be tempted to banish all external res from the belief canonically attributed to one who asserts that Fabc... But McGinn’s contextual theory is itself neutral as between narrow and wider functionalisms. It is consistent with the asserter having
beliefs are the public expressions employed in his act of assertion. And if one can have thoughts about expressions, then, given suitable perceptual and manipulative contact, one can have beliefs about cars, people, etc. I hope that where an attitude is a particular res, that res is a first-order constituent of the second-order state which realises that attitude.

How should the A-BG be modified to incorporate the above? We could avoid the need for extensive re-working by building in a blanket condition, "Cxt", that x be in suitable perceptual or manipulative contact with all the referents of any singular terms in the token t, thus:

$$(x)(p)(Nx \rightarrow (x \text{ believes } p \leftrightarrow (\exists t)(Axt \land Cxt \land s(t)=s(p) \land (\exists s)(P_{ts} \rightarrow P_{ps})))$$

But at the price of severely restricting the application of our operational test.

In general, the asserter will have beliefs about some of the referents but not all. Suppose x asserts "Fabcd", and suppose he has suitable perceptual and manipulative contact with a and c, but not with b or with d. Then the appropriate belief report is

$$x \text{ believes } [(\exists y)(R"b"x \land R"d"y \land Faxcy)]$$

further modification being needed, if any "a" to "d" are indexicals, due to the different contexts of believer and belief-reporter. Clearly, extensive re-working of the A-BG is needed to cover cases where t may or may not contain singular terms, and x may or may not have the required perceptual or manipulative contact with any one of these various referents.
Footnotes to Chapter Two

1. Loar (1)
2. Ibid. Ch 4 and Ch 7
3. Ibid. p62
4. Ibid. p72
5. I am using "instrumentally" loosely, as an umbrella term for semantic accounts which range from treating the apparent theoretical entities as logical fictions or constructs, to van Fraassen's (1) "constructive empiricism", in which one who asserts a theory commits himself only to the truth of its implied observation sentences, and not to the unobservable entities and facts the theory appears to postulate to explain the observable. What all these views have in common with the sentential index account of psychological explanations, is that applying the theory does not commit one to the total putative ontological package, entities and facts.
6. Loar (1) p151
7. To transfer Nagel's famous phrase, Nagal (1), to Tom's cognitive and volitional states.
8. For these reasons, an alien intelligence, the Martian anthropologist, forearmed with the conceptual resources of Martian, and to whom a different selection of first-order physical facts about Tom are salient and readily available, may be unable to see us as psychological subjects, may think there is nothing it is like to be human. Or, perhaps, the natural first-order facts which are salient to him about Tom and his fellows 'stack up' into a second-order system whose states can be indexed by two systems of Martian sentential indices, two different Martian psychological histories of Tom, ascribing different Martian conceptual contents to Tom's second-order states. If so, my criterion requires the conceptual contents of these indices be taken instrumentally by the Martian theorist. Or a more bizarre possibility is that the Martian, with his different perspective on the first-order facts about Tom, and different conceptual repertoire, may
see those first-order facts as stacking up into a
different second-order system, a system whose states are
uniquely indexed in Martian sentences. If my criterion
requires me to really apply the conceptual contents of
the sentential indices of my unique psychological
history of Tom to Tom's internal states, does it not
also require the Martian to do likewise? If so, then Tom's
physical body is the locus of both a human-type
psychological history and the locus of a Martian-type
psychological history. There is no incompatibility at
the physical level, because those first and second order
physical facts about Tom which are salient to us humans
are consistent with the different first and second order
physical facts about Tom which are salient to Martians.
And there is no incompatibility at the psychological
level, provided the set of second-order physical facts
which realise the human conceptual contents is different
from the set of second-order physical facts which realise
the Martian conceptual contents.

9 Fodor J(3):"Tom Swift and his procedural grandmother"
10 Peacocke (2) Ch3
11 Ibid. pp78 - 86
12 Peacocke (2) p83 and Evans (1)
13 Peacocke (2) p83

Note the way Peacocke aligns "kinds of sentences" with
"kinds of actions" in the first sentence of the passage
just quoted, without remarking that Quine and Evans take
sentences to be sorted into types prior to translation,
but behaviour is not sorted into action kinds prior to
interpretation.

15 Fodor J (3):"The present status of the innateness controversy"
16 Fodor (1) passim
17 Peacocke (2) Ch8
18 Davidson (3):"On saying that"
19 Davidson (3) p xiv and p53
20 Ibid. p94
If we try, instead, to interpret "Tom believes-true p", where p is some eternal sentence, as a conditional about Tom's behaviour, to the effect that Tom would assent to p, if ..., then we run into trouble. Eternal sentences in Quine's regimented language may be very syntactically complex, and Tom quite stupid. If we build into the antecedent of the behavioural conditional that Tom is given a course of intelligence pills and taught Quinese, then we may expect lots of his beliefs to change. And, for Quine, we cannot isolate, in any principled way, his target belief that p from these shifts in his web of belief, thus destroying the claim of the conditional to explicate unreformed Tom's belief that p.

If there is a language of thought, one may postulate that this distinction between semantic and syntactic inconsistency carries over to the inscriptions of MENTAL within the subject's head. Then a subject holds syntactically inconsistent beliefs iff P, Q \not\vdash *, where P and Q are two of his inscriptions in MENTAL, and he holds semantically inconsistent beliefs iff P, Q \vdash *. We expect that if his beliefs are syntactically inconsistent then they are semantically inconsistent also, but we have no good reason to expect the reverse. And a subject's logical acumen is measured by his ability to avoid syntactic inconsistency.

There remains, of course, the possibility that Pierre's total behaviour fragments into two mutually exclusive subsets, one of which leads us to index his beliefs by "London is pretty" and the other of which leads us to
index his beliefs by "London is not pretty". But such is not the present case: here we are considering one action whose intentional explanation seems to involve syntactically inconsistent indices.

31 McGinn (1)
Chapter Three: On the alleged indeterminacy of conceptual content

Section (i): Introduction

In the last chapter I assumed that the sentences by which we index our own and fellow language speakers' cognitive and volitional states have determinate conceptual contents. In this chapter I will do what I can to defend that assumption. It is threatened by the arguments of Quine, Putnam and Davidson for the inscrutability of reference, because, if the reference of a given utterance in a given context is inscrutable, this must be because the conceptual content is indeterminate - conceptual content plus context determining reference, if anything does.

Some arguments purport to find inscrutability at the level of sub-sentential words and phrases, but not at the level of whole sentences. Others find inscrutability at the level of whole sentences too. It would not be sufficient for my criterion of psychological realism if sentential content only were determinate, being measured, perhaps, by fine-grained truth values across possible worlds, whilst indeterminacy infects the conceptual content of sub-sentential phrases. It would not be sufficient because, as we have seen in chapter two appendix B, psychological explanations involve indexing the subject's states by open sentences in the style of

$$(\exists x)(\text{Tom believes of } x \text{ that } [\text{London}x \land \text{Pretty}x])$$
where the index "Londonx₁ & Prettyx₁" contains a quasi-bound variable, since it is not assigned any specific value. So the conceptual content of the index is that of a phrase rather than that of a whole sentence, and if the conceptual content is to be applied to the internal state of Tom so indexed, the index must have a determinate conceptual content. Hence, I shall be attacking arguments for sub-sentential indeterminacy as well as arguments for sentential indeterminacy.

Arguments for inscrutability and indeterminacy are legion, and the resulting literature enormous. I do not claim to have read more than a fraction of it, and shall not comment on all that I have read. Clearly, rejecting certain arguments for indeterminacy risks neglecting better ones, and my enterprise does not amount to a positive argument for determinacy. However, that family of arguments sired by Quine's gavagai argue for a surprising indeterminacy: no-one thinks that his use of "rabbit" is indeterminate between the concepts of rabbithood, undetached rabbit-part, rabbit feature, etc. So far as these arguments are concerned I shall take the burden of proof to lie with those who argue for indeterminacy. For conceptual content is a conventional non-natural property of marks and sounds, and a convention is a correlation between utterance types and something else, which correlation holds because it is known by us to hold, we being parties to the convention. Now it may be that
some of the implications of our conventions are unknown to us, and it may be that we cannot know all the implications of our conventions; perhaps, even, there are some specific implications we cannot know. But the basic, meaning-bestowing conventions are known to us, and the indeterminacies mooted in the arguments are on such a scale that, it seems, they couldn't remain unrecognised by us once pointed out. Hence my polemical aim will merely be to show that the 'gavagai' argument and its permutational progeny developed by Davidson and Putnam do not establish inscrutability of reference. And, I think, this is all the defence my claim of coarse grained referential scrutabilty and conceptual determinacy requires.

Quine's 'Duhemian' argument is a fish from a different kettle. It does not claim an obviously surprising indeterminacy. A proponent of the Duhemian argument may confine his indeterminacy to the higher reaches of theory, where the connection between language and the world is indirect, being established on the back of manifest use of language in response to the observable world. Merely showing that the Duhemian argument does not establish indeterminacy results in a draw: there is no prima facie presumption that our indirect ways of linking theoretical language with the world result in a unique linkage.

A propos of the 'gavagai' argument and its permutational offspring, I eschew the "myth of the museum" for the same reasons as does Quine: if the physical facts fail to determine a unique extension and conceptual content, then
additional mental facts cannot "bridge the gap", because psychological facts are realised and constituted by physical facts. An alleged psychological difference between Tom, a closet rabbit-referer, and Dick, a closet undetached rabbit-parter, and Harry..., would have no constituting or realising physical basis. It is no answer to reply that if Tom, Dick and Harry do not differ in any relevant physical features, and if their usage of "rabbit" leaves its extension undetermined, but Tom is a closet rabbit-referer, then so must be Dick and Harry. Certainly, this would formally preserve the supervenience of the mental upon the physical, but at the cost of making that supervenience magical: if the physical facts equally accord with Tom, Dick and Harry being undetached rabbit-parters, and equally accord with them being rabbit featureurs, etc., then how those physical facts constitute and realise the boys all being the first and not the second nor the third is utterly incomprehensible.

Section (ii): Putnam on the indeterminacy of reference

Let me first present Putnam's argument in an attenuated form and show why it will not work in this attenuated form. By doing so I hope to clear the clutter and expose the real core of Putnam's argument.

Putnam establishes a lemma:  

no view which only fixes the truth values of whole sentences can fix reference, even if it specifies the truth values for sentences in every possible world
Actually this lemma is established only for a regimented satisfactional-functional language. Elsewhere Putnam cites the Lowenheim-Skolem theorem in its support, which Putnam generalises informally from first-order theories to the sentences which index psychological states\(^6\).

Let us grant this lemma applied to the sentences which index psychological states, although those indices are not exclusively first order, and whether they are satisfactional functional is a moot point.

Let us suppose, then, that more than one assignment of referents to the singular terms and extensions to the primitive predicates of our indexing language L yield the same truth conditions for the sentences. Putnam develops an example: under one interpretation the extension of the word "cat" in any possible world are the cats in that world, and, under a second interpretation of L, the extension of "cat" is the set of cats in some possible worlds and the set of cherries in other possible worlds. And, under the second interpretation, the extension of "cat" is the set of cherries in our actual world. Putnam introduces the term "cat*" explicitly to have the extension (and intension) assigned to our term "cat" by the second interpretation: i.e., the set of cherries in the actual and in some other possible worlds, and the set of cats in some non-actual possible worlds.

Putnam raises the question whether there is something else, over and above truth conditions, which can fix reference. An imagined interlocutor proposes, as the
additional determining factor, the causal relations involved in perception and action. In effect, the interloctor is protesting that, whereas truth values and therefore inferential inter-relations between indices do not determine references and extensions; nonetheless, references and extensions are determined by the input and output conditions of the functional psychological theory. It is because, inter alia, the belief state indexed by "The cat is on the mat" is differentially sensitive to whether or not a cat is on a mat in the causal vicinity of the psychological subject, but not to the facts about cherries, if any, in his neck of the woods, that "cat" refers to cats and not to cats*. In short, proximity to cats causally triggers "cat"-indexed belief states, which causally trigger, in turn, kickings of cats. But there are no comparable salient causal relations between these psychological states and cherries.

However, Putnam dismisses the protest as follows:

Unfortunately, one can reinterpret "sees" (say, as sees*) so that the two sentences (1)John (or whoever) sees a cat, and (2)John sees* a cat*, will have the same truth value in every possible world ... So whenever a person sees a cat, he is seeing* a cat*; the experience we typically have when we see a cat is the experience we typically have when we see* a cat*, and so on. Similarly, we can reinterpret "inspects" and "tells" so that, when [John] inspects a cat, [he] is inspecting* a cat*, and when [he] tells something is a cat, [he] is telling* that it is a cat* .......

The point is that the fact[sic] that one can build a machine to inspect things and tell if they are cats differentiates cats from cats* if one can be sure "inspect" and "tell" refer to inspecting and telling, and it is no easier to say how the reference of these words are fixed than to say how the reference of "cat" is fixed.
Thus, it seems, input and output conditions fail to determine the referents and extensions of the indices.

Putnam proves his lemma in a metalanguage $L'$. In $L'$, the object language $L$ is specified, and extensions for the primitive terms are supposed given, and then a permutation is specified, or more accurately a class of such permutations, which preserves the truth values of sentences of $L$ in all possible worlds. Then the protester says

A human is a cat detector, but not a cat* detector ...(l)

This remark is part of a wider causal-cum-biological theory of humans, and is intended to favour the first interpretation of $L$ over the second. Clearly, if this remark is intended to select one rather than the other interpretation then it had better be framed in the same language as those interpretations, the metalanguage $L'$. So when Putnam ripostes that (1) has itself more than one interpretation, he is relying on a parallel proof of his lemma applied to $L'$, that proof being conducted in $L''$. The semantic ascent in evident in such passages as$^9$:

If 'refers' can be defined in terms of some causal predicate or predicates in the metalanguage of our theory, then, since each model of the object language extends in an obvious way to a model of the metalanguage, it will turn out that, in each model $M$, reference, is definable in terms of causes,; but, unless the word 'causes' (or whatever the causal predicate might be) is already glued to one definite relation with metaphysical glue, this does not fix a determinate extension for 'refers' at all.

The argument, as we have it so far, fails to establish inscrutability. At any level in the hierarchy of languages and their metalanguages, he may propose alternative interpretations, but, at that level, causal theory selects
one of those interpretations. No matter that the meta-
language, and hence the causal theory framed in it, have
various interpretations, provided that, at whatever level
rival interpretations are formulated then, at that level,
causal theories select one interpretation. At no stage
in the hierarchy do we reach a level at which the other
theories at that level fail to select one of the rival
interpretations at that level, so at no stage are we
warranted to claim that interpretation of the lower stages
is indeterminate.

Imagine a linguist set the task of constructing a
semantic theory for an object language L. To do so he
will need a metalanguage L' to express his theory, and to
express whatever background physical and biological
theories he considers relevant. Imagine a second linguist
set the task of constructing a semantic theory for L',
equipped with a metalanguage L'' in which to express his
theory and other relevant background theories. And so on.
Each of these linguists will acknowledge that, given one
interpretation, they can generate other interpretations
salva veritate of their object language in their meta-
language. But none need regard this as showing that his
object language is referentially indeterminate, since
background theories will select one interpretation, we
suppose. What difference does it make to their conclusions
when each realises his place in the sequence and sees that
the others all give the same answers as himself?

Putnam has assumed that reference is indeterminate in
the object language, and that this indeterminacy rises upwards through the hierarchy so as to render the causal theory of reference in the metalanguage itself indeterminate. But this is to beg the question. Putnam has admitted, at least for the sake of argument, that other warranted theories in the metalanguage may favour one interpretation over the others. But that is to admit that he has not yet shown that multiple interpretations salva veritate are sufficient for indeterminacy of reference and extension. He then seeks to re-establish indeterminacy by pointing out that these metalinguistic theories themselves have multiple interpretations salva veritate. But the fallacy is now obvious: he is not entitled to assume that multiple interpretations salva veritate are sufficient to establish indeterminacy. There is a lacuna in this argument extracted from Putnam: the protester is entitled to cheerfully agree that the metalinguistic theories which select, at any level, one of the proffered interpretations are themselves open to multiple interpretations salva veritate, since, the protester maintains, the argument, at this stage, is working under the assumption that interpretation salva veritate is not sufficient to establish indeterminacy.

The passage last quoted from Putnam suggests remounting the argument along a dimension of metalanguages 'orthogonal' to the first. Suppose an object, meta-, metameta-, etc. language hierarchy as above, and suppose that each language is multiply interpreted in the one above but there is an accepted causal theory in the language above which favours one of the interpretations. We can now variously interpret
the object language, *salva veritate*, in a *new* metalanguage $L'_1$ and allow each interpretation of $L$ in $L'_1$ in turn to determine the interpretation of $L', L'', ...$ via the favoured interpretation of $L$ in $L'$, and of $L'$ in $L''$, etc. Now, although the relations between $L$, $L'$, $L''$, ... are fixed, one mapping of the sentences of $L$ onto $L'$ and of $L'$ onto $L''$ and etc. being salient, the favoured line through $L, L', L'', ...$ is variously interpreted in $L'_1$. So the reference of the terms of $L$ (and therefore of $L', L'', ...$) is inscrutable.

The answer to this remounting of the argument is the same as to the original argument. The mere possibility of various interpretations *salva veritate* of $L, L', L'', ...$ in $L'_1$ is not sufficient to establish inscrutability. $L'_1$ is itself an interpreted language with a causal and other theories. And those other theories favour one interpretation of $L$ in $L'_1$, and therefore derivatively one interpretation of $L', L'', ...$. Nothing has been gained.

Now we come to the central core of Putnam's argument. I have omitted, so far, a central premiss in Putnam's argument. That premiss is: reference (extension) is *not* a natural relation, but is an *intentional* relation. So, he thinks, granted that a set of sentences, or internal representational states, has more than one interpretation *salva veritate*, NO purely causal relations, taking causality to be a naturalistic concept, linking those sentences or representations to the world, no matter how salient, *can determine* the intended interpretation of those sentences or representations. No matter that
assigning the set of cats as the extension of "cats", rather than the set of cherries in some possible worlds and cats in others, 'tracks' the causal conditions under which "cat"-utterances are produced in the various possible worlds. Nothing purely naturalistic can determine an intended interpretation.

I hope it is clear that this argument is best disentangled from the earlier one employing hierarchies of object and metalanguages. The present claim is that causal facts cannot determine reference, being facts of the wrong kind. Of course it follows that reference is indeterminate in all languages, object and metalanguages. But the argument is not best put as the claim that constraints upon interpretative theories (sets of sentences) do not uniquely favour one over the others. Rather, the argument is that given the truth conditions of the sentences of the object language, and given the causal facts about sentence usage, those facts cannot select one interpretation salva veritate, and so reference is indeterminate, there being nothing else to select one interpretation.

In Putnam's own words, he puts forward the suggestion

Reference is not fixed by anything psychological, anything "inside the head". But that is no problem: why can't reference be fixed by something non-psychological? The something non-psychological in question being causal relations, external to the head, between utterances and features of the world. But Putnam immediately replies:

The answer, quite simply, is that the idea that the 'non-psychological' fixes reference - i.e., that nature itself determines what our words stand for - is totally unintelligible.
Putnam’s claim is that physical facts cannot determine relations of reference or extension. We need to separate this claim from two other claims. Firstly, we have agreed that psychological facts, including semantic facts if such there be, cannot be reduced to or defined by physical facts. Thus we can agree with Putnam that, if R is some physical relation, and "R is reference", then that would be "a metaphysically unexplainable fact, a kind of primitive, surd, metaphysical truth". The view that some physical correspondence between representations and the world is reference "amounts to a magical theory of reference". But the view to be examined is not that reference is a physical relation, but that reference is determined by a physical relation.

Secondly, we need to separate the claim that reference is not determined by physical facts from narrow functionalism: from the view that nothing external to the head determines any psychological fact. What seems to determine that the extension of "cat" is cats rather than cats* is that 'cat' utterances are triggered by the causal proximity of cats rather than cherries. But these causal facts are a fortiori irrelevant if nothing external to the head is allowed to determine any psychological fact. Here is a passage where Putnam argues from narrow functionalism to the irrelevance of external causal relations:

if 'mentalese' is to be a vehicle for describing the external world, then the various predicate letters must have extensions which are sets of external things (or sets of n-tuples of external things). But if the way 'mentalese' is 'understood' by the
deep structures in the brain that compute, record, etc. in this 'language' is via what Artificial Intelligence people call 'procedural semantics' - i.e., if the brain's program for using 'mentalese' comprises its entire 'understanding' of 'mentalese', where the program for using 'mentalese', like any program, refers only to what is inside the computer - then how do extensions ever come into the picture at all? ... If thinking is ultimately done in 'mentalese', then no concept will have a determinate extension. Or so it seems.

And a few pages later:

Nor do 'causal theories of reference', etc. help. Basically, trying to get out of this predicament by these means is hoping that the world will pick one definite extension for each of our terms even if we can't. We interpret our languages or nothing does.

The issue between narrow and wider functionalisms, and the relation between machine-codes, procedural semantics and the concepts we grasp, are different issues from the claim that physical facts cannot determine psychological, including semantic, facts. It is as much or as little a mystery as to how the physical facts within our skulls determine machine code conceptual and semantic content as it is how wider physical facts determine conceptual and semantic content about cats. It is an additional thesis, one not up for discussion in this chapter, that the conceptual content of utterances are to be defined by the concepts of the machine code. Only with that additional premiss do we reach the conclusion that the physical facts which determine reference, if any do, must be within the skull of the referer. If the world within the skull can pick the mentalese concepts of procedural semantics, then the wider world may pick the concepts we grasp in our utterances, provided the latter concepts are not defined
by the former.

We are now ready to present what is, I think, Putnam's best argument for the indeterminacy of reference. We suppose that the truth conditions of a person's utterances are determined. As naturalists we suppose that they are determined by the physical facts, be those narrow or wide physical facts. Now the references of the sub-sentential parts are determined, so far as they are determined, only via the referential intentions of the utterer. However, the truth conditions of his intentions, of his utterances and of all his propositional attitudes taken together do not determine a unique set of referents, there being more than one interpretation salva veritate of his utterances. Indeed, we may suppose that under one interpretation the referential relations are 'aligned' (coextensional) with certain causal relations, whilst under the other interpretations the physical relations between a term and its extension are ad hoc. But although physical facts can and do determine the truth conditions of his intentions, these remarkable physical facts cannot determine the intended interpretation. Physical facts can determine the truth conditions of beliefs, intentions, utterances, etc., but their potency is thereby exhausted. If sub-sentential reference is determined, it is determined by those truth conditions. Therefore it is not determined.

Why is sub-sentential conceptual content and reference determined, if it is determined, only by the truth conditions of propositional attitudes? Because all psychological
explanation is governed by the master concept of rational behaviour; the explanation shows how the behaviour was rational. And rationality is a matter of consistency of attitudes. But the attitudes, including intentions to refer, are as mutually consistent under any one interpretation salva veritate as under any other. The causal salience of one interpretation relative to the others, or any other physical features within or without the skull, do not make the behaviour under that interpretation any more rational. So physical facts cannot determine one interpretation rather than the others.

The crucial question raised by Putnam's argument is: Can the physical facts determine the truth conditions of representational states without thereby fixing determinate sub-sentential conceptual and semantic contents? To tackle this question it will help to look first at the work of Davidson.

Section (iii): Davidson's sub-sentential instrumentalism

Davidson adds four interesting theses to Quine's indeterminacy of radical translation and consequent inscrutability of reference: (i)translation manuals should be modelled on Tarski's definition of truth, (ii)only the T-theorems of the theory are empirically testable against native behaviour, (iii)only the semantic concept of truth is to be interpreted realistically, that being the only semantic concept which occurs in the T-theorems, and (iii)the sub-sentential apparatus of the Tarskian semantics whereby referents are assigned to singular terms.
and denotations to primitive predicates is to be interpreted instrumentally. Thus the Tarskian theory\(^{15}\)

will, of course, contain a recursion on a concept like satisfaction or reference. But these notions we must take as theoretical constructs whose function is exhausted in stating the truth conditions for sentences... None of this apparatus is open to direct confrontation with the evidence.

And\(^{15}\)

The theory gives up reference, then, as part of the cost of going empirical... We don't need the concept of reference; neither do we need reference itself, whatever that might be... Reference drops out. It plays no essential role in explaining the relation between language and reality.

And the reason for this instrumental downgrading of sub-sentential semantics compared to the reality of sentential semantics is the inscrutability of reference relative to fixed truth conditions\(^{16}\):

if there is one way of assigning entities to expressions (a way of characterising 'satisfaction') that yields acceptable results with respect to the truth conditions of sentences, there will be endless other ways that do so as well.

So far, nothing essentially different from Putnam.

However, Davidson's picture allows an important theoretical position for the real salient causal relations between "cat"-utterances and cats, in contrast to the ad hoc real physical relations between "cat"-utterances and cats*. Davidson originally introduced truth theories as a model to explain how we finite creatures can learn and understand a language, how we can compute and grasp the truth conditions of any of an indefinitely large number of sentences we have never heard before. Clearly we cannot take the posit of an internal metalanguage
relistically in pursuit of this explanatory goal, on pain of lauching the explanation upon an infinite regress. But if the reference (satisfaction) relations are not to be interpreted realistically either, then, seemingly, nothing of the Tarskian model survives to explain our grasp of our mother tongue, and the original problem is unsolved.

Davidson can have it both ways. The salient causal relations between, for example, our "cat"-utterances and cats, which do not hold between, for example, our "cat"-utterances and cats*, are real and do explain our marvelous ability to grasp the truth conditions of new sentences. These causal relations between names, predicates and entities in the world explain our productions of utterances containing those names and predicates. And also, although one of the theoretically posited satisfaction relations is coextensive with this causal relation, this does not give that satisfaction relation a special status in the semantic theory: it is still, like any of its alternatives to be interpreted instrumentally.

Why does the coextensiveness of one particular satisfaction relation with the salient causal relation not make that satisfaction relation real, determined by the causal relation? Because, thinks Davidson, the empirical test of a truth theory lies in the T-theorems, which employ only the semantic concept of truth applied to whole sentences. And because truth conditions under-determine the satisfaction relation. So even though there is this
extraordinary causal salience coextensive with one and only
one theoretical satisfaction relation, and even though the
language users can recognise this unique coextensiveness,
it being public and observable, that is not sufficient to
make that satisfaction relation real - because there is no
way they can show in their usage of whole sentences that
they have selected it as the satisfaction relation rather
than one of the others.

Consider, in this light, the following passage¹⁸:-

Perhaps someone ... will be tempted to say, 'But at
least the speaker knows what he is referring to'.
One should stand firm against this thought. The
semantic features of language are public features.
What no one can, in the nature of the case, figure
out from the totality of the relevant evidence
cannot be part of meaning. And since every speaker
must, in some dim sense at least, know this, he
cannot even intend to use his words with a unique
reference, for he knows that there is no way for
his words to convey this reference to another.

Remember that the salient causal relations are themselves
public and readily observable: "cat"-utterances being
triggered by cats and not by cherries. So the crux upon
which Davidson's argument turns is that there is no way
a speaker can indicate that he intends this relation rather
than another, and this because the interpretation of his
utterances rests upon, and only upon, assigning the intended
truth conditions to his sentences.

Truth is the only semantic concept employed in Tarski's
T-theorems because the sentences of the object language
are context free. Although the use of proper names is not
really context free - there is Aristotle the philosopher
and Aristotle the tycoon - we can unbind them by treating
the tokens as tokens of different word-types, distinguished by indices in the metalanguage - "Aristotle₁" and "Aristotle₂". But indexicals are a different matter: there is no fixed list of the referents of "that", rather one must grasp a certain function from context to referent. Grasping the particular function associated with each indexical term is part of understanding the language. Now Davidson is able to deal with some indexical terms without introducing any new semantic concept into the T-theorems. He deals with "I" and "now" by constructing a truth theory which yields, for example, the theorem¹⁹: -

"I am tired" is true as spoken by x at t iff x is tired at t without introducing any sub-sentential semantic relation into the theorem.

Such manoeuvres only postpone the real problem: pronouns like "we", "you", "she", "it"; demonstratives like "that" where (i) there is no finite list of referents as there was with "Aristotle", and (ii) an operational test is needed to pick out the referent, unlike "I" or "now". Various writers²⁰ have sought to extend Davidson's theory of meaning to encompass such demonstratives. But in all cases which I have seen, the extensions yield theorems which employ a semantic relation linking token indexical to referent, as well as the semantic relation of truth. Since these theorems are the 'observation sentences' of the theory of meaning, the field linguist will need an operational test for this new semantic relation in order to test the theorems against native usage.
Let Barry Taylor's theory stand for all. His theorems employ a function Dem (Demonstrates) which maps an utterance u of "that" at a time t onto an object. The T-theorem for the Native sentence "Red(\text{that})" is:

\[
\text{True}(\text{Red(\text{that})},u,t) \iff (\exists x)(x=(\forall y)\text{Dem}(u,t,y) \& \text{Red } x)
\]

The introduction of "Dem(u,t,y)" into the observation sentences by which a truth theory is empirically tested would not matter to Davidson if "Dem(u,t,y)" reported a physical, causal relation. But it does not. English speakers may be causally disposed to utter "horse" in the causal vicinity of horses, but the relation between an utterance of "that horse" and one particular equine occupant of the paddock is conventional: the rule is to project a line from the knuckle to the fingertip of my prominently displayed finger until it intersects a horse which is visible to you, the audience. Other cultures do it differently.

Conventional relations are real relations which hold because parties to the convention know that they hold. Davidson admits the need for such conventional relations, but seems to think that they are somehow ancilliary to the theory of meaning. Thus he writes:

The verification of instances of the T-sentences, or rather their surrogates in a theory reativised to speakers and times, remains respectably empirical. No doubt some pragmatic concept of demonstration as between speakers, times, and objects will come into play. But the concept is one we may hope to explain without appeal to notions like truth, meaning, synonymy and translation.
Davidson is admitting, I think, that some conventional, semantic relation linking demonstratives to their referents is real and is needed by a theory of meaning, since it occurs in the empirically tested 'observation' sentences of a Tarskian theory of meaning. But, in calling it a pragmatic concept, he is indicating that it is not to be identified with the theoretical concept of reference (satisfaction), which is still to be interpreted instrumentally.

However, we can now turn the screw on Davidson. If the radical interpreter has to adopt a convention relating demonstratives to entities then so too has the Native. For Davidson has claimed that the semantic features of language are public; the Native is not in a position which is, in principle, epistemologically or ontologically different from that of the interpreter. Thus the interpreter must suppose that the Native too has adopted some convention relating sub-sentential parts to entities. The question arises whether the interpreter can discover which convention the Native has adopted? Is the Native's convention scrutable?

Of course the Native's convention is scrutable. If one convention seems obvious to the interpreter, whilst others seem bizarre and perverse, then he has his warrant to suppose that the Native too adopts that convention, so far as the Native seems a regular chap, giving no evidence of a bizarre or perverse personality.

It might seem that there is a fatal circularity here: the Native's behaviour is that of a regular guy iff he has
adopted the same convention as you. There is no circularity.

Davidson had claimed that there was no fact of the matter, even in your own case, linking sub-sentential parts to particular lumps of the world in some conventional relation. But we have seen that you have to adopt just such a convention to interpret utterances. And one convention stands out as natural and obvious to you: you just cannot work the alternative conventions whereby you would demonstrate* cats*, believe that you were seeing* a cat*, intend* to kick* a cat*, etc. But you are a creature of the same physical kind as the Native, so naturalism requires that what is psychologically evident to you is so to him also.

Of course, the theorist may, if he wants to, introduce a theoretical relation of satisfaction, whose only function in the theory is to generate truth conditions and which is not to be interpreted realistically. The point of doing so may indeed be to demonstrate that rationality, conceived of as consistency of attitudes, does not determine reference. But to do so would not be to deny the reality and necessity of a convention of reference, linking sub-sentential parts to the world. There is a fact of the referential matter because rationality, conceived of as the mutual consistency of attitudes, presupposes that there really is a convention of reference.

My reader will have been reminded of the traditional argument by analogy for other minds. That argument was, roughly, that I know in my own case that my physical
behaviour is accompanied by various mental goings on, I observe similar patterns of behaviour in others, so I infer, by analogy, similar mental events in their cases too. My argument differs in the following respects:

(i) I do not suppose that the radical interpreter adopts some private convention, only that he adopts some convention, satisfying whatever necessary conditions as may be. I have entered no opinion on whether or not one may follow a rule in private. (ii) It is not a risky inference that the Native too operates a convention, that is given. So (iii) the issue is which convention does he use, is it the same as ours? I have answered this question from the standpoint of naturalism: what seems obvious to me must also seem obvious to him, we being creatures of the same physical kind. But one could also argue that it is a necessary condition of any convention that like behaviour manifests the same rule to humans.

My answer to Davidson's sub-sentential semantic instrumentalism is that a theory of interpretation must employ a real convention relating sentence parts to entities, and so therefore must the Native. Hence the only question which remains is: Does he adopt the same convention? And, given physical parity, the answer is 'Yes', either on the ground of naturalism, or by an argument from the necessary conditions for a conventional practice.

What, then, of Putnam's argument of the last section? Putnam argued that psychological explanation is constrained
by rationality, conceived as requiring the mutual consistency of attitudes. Supposing the physical facts determine the truth conditions of attitudes, they will thereby determine a mutually consistent set of attitudes. But the truth conditions do not determine a set of referential relations. So the physical facts, in determining psychological facts do not determine referential relations.

But now we see that, although truth conditions do not determine reference (there is no route from truth conditions to reference), physical facts determine truth conditions only by determining some referential relations (the route to truth conditions is via reference). And, not surprisingly, the salient causal relations between "cat"-utterances and cats, rather than between "cat"-utterances and cats*, determine that "cat" refers to cats, and not cats*.

The radical interpreter may, for his own purposes, introduce a theoretical relation, satisfaction, subject only to the constraint that it deliver up the right truth conditions, and therefore to be interpreted instrumentally. Also, the radical interpreter may introduce into his theory defined relations along the lines

\[ \text{Dem}^*(x,y,t) \text{ iff}(\text{Cat}^y \land (\exists z)(\text{Dem}(x,z,t) \land \text{Cat}z \land F(z) = y)) \]

where \( F(y,z) \) is some one-one function which takes cats onto cats*. Then, whenever a Native demonstrates a cat he demonstrates* a cat*, and demonstrates** a cat**, and ..., since to do the one is to do the others. Demonstrating* a cat*, for example, is as real as demonstrating a cat, because demonstrating* a cat* is demonstrating a cat.
And the radical interpreter can, of course define \( \text{Dem}(x,y,z) \) in terms of, say, \( \text{Dem}^* \) as follows:

\[
\text{Dem}(x,y,z) \iff (\text{Cat}y \& (\exists z)(\text{Dem}^*(x,z,t) \& \text{Cat}^*z \& F^{-1}(z)=y))
\]

However, the radical interpreter cannot be neutral about which, if any, of \( \text{Dem} \), \( \text{Dem}^* \), \( \text{Dem}^{**} \), ... is ontologically basic. He sees a cat and he kicks a cat, and he can take himself to thereby be seeing* a cat* and thereby kicking* a cat*/only by taking "seeing* a cat*" and "kicking* a cat*" to be definitionally equivalent to "seeing a cat" and "kicking a cat". On the other hand, he can take himself to be seeing a cat and kicking a cat without any knowledge of the definitional relations between "seeing" and "seeing*" or between "cat" and "cat*", or between "kicking" and "kicking*". And further, he cannot find this partiality in himself and not suppose the same partiality in the Native.

Suppose a theorist puts forward a disjunction of truth 'theories'

\[ T \lor T^* \lor ... \]

which will then yield disjunctive theorems, e.g.

\[
\text{True("Cat that",u,t)} \leftrightarrow (\exists x)(x=\text{Dem}(u,t) \& \text{Cat}x)
\]

\[ \lor \]

\[
\text{True("Cat that",u,t)} \leftrightarrow (\exists x)(x=\text{Dem}^*(u,t) \& \text{Cat}^*x)
\]

\[ \lor \]

to be tested against native behaviour. Would this be an empirically testable truth theory which incorporates referential inscrutability? The instrumental nature of the relations \( \text{Dem} \), \( \text{Dem}^* \), ... is evident from the fact that we can confirm the disjunction but not any individual disjunct?
Whilst the reality of the truth property is evident from the fact that it is ascribed in each disjunct? Unfortunately this is impossible: a disjunction is observationally warranted only if one (or more) disjunct is observationally warranted. But then the Dem (or Dem*, or ...) relation of that disjunct has the same claim to be taken realistically as the truth property. Verifying the truth theory and treating the truth property realistically presupposes treating one or more of the demonstration relations realistically also.

Nor does it help if the theorist puts forward a conjunction of truth 'theories'

\[ T \land T^* \land \ldots \]

which will then yield conjunctive theorems, e.g.

\[ \text{True("Eat that",u,t) } \leftrightarrow (\exists x)(x=\text{Dem}(u,t) \land \text{Cat}x) \]

\[ \land \text{True("Cat that",u,t) } \leftrightarrow (\exists x)(x=\text{Dem}^*(u,t) \land \text{Cat}^*x) \]

\[ \land \text{ etc.} \]

A sentence of the form \( P \land Q \land \ldots \) is observationally warranted only if \( P \) is observationally warranted, and \( Q \) is observationally warranted, and \( \ldots \) Further, the conjuncts, being observation sentences of the interpretative theory, are to be interpreted realistically. So we can present our theorist with a dilemma: Do the various conjuncts report the same sub-sentential relations, "Dem", "Dem*", ... being interdefinable? If so, there is a matter of fact as to how parts of sentences are related to entities. And if there is such a fact of the matter, we can, in principle, determine it - to the limits of empirical defeasibility. Or, to take
the other horn, does each conjunct present a distinct sub-sentential relation, such that none is primitive relative to the others, but all are ontologically and epistemically on a par? But I am not a creature who can operate an indefinitely large plethora of conventional relations simultaneously, and nor are you I think. It may be an insuperable empirical problem for a Martian to discover which convention humans operate, given that any of them would yield the same truth conditions. But even the Martian, if his conception of a theory of meaning is at all like ours, knows that we humans operate some convention, if our utterances do have determinate truth conditions. But it is no empirical problem for me to know which sub-sentential convention I operate. So it is no problem for me to know which convention you operate either.

The onus of proof was on Putnam and Davidson to show that reference is inscrutable, there being no fact of the matter. But modelling theories of meaning on truth theories has not shown this. They have shown that truth conditions do not determine reference. But not that physical facts can determine truth conditions without doing so via referential relations. They have not shown that there can be a fact of the matter about truth conditions without there being a fact of the matter about referential conventions.
Section (iv): Quine's Duhemian argument

Quine argues that high-level theoretical statements are semantically indeterminate.22 If so, given my Fregean view of concepts as the determinants of reference, their conceptual contents must be equally indeterminate. However, since the argument concerns high-level theoretical statements, rather than our bread-and-butter verbal responses to the world, the claims of indeterminacy and inscrutability lack the Quixotic air of the earlier arguments. So I make no presumption about the burden of proof: spinning the web of theory may well generate unnoticed and unintended indeterminacy, as it has, on occasion, generated unnoticed and unintended paradox.

Duhem observed that there is always more than one way of squaring the totality of theory with a recalcitrant experimental result. Building on this underdetermination of theory by experiment and observation, Quine assumes some perhaps arbitrary line drawn between the native observation sentences and the native theoretical sentences, and he concedes that we have determinate translations of all native observation sentences, and Quine argues that the translation of native theoretical sentences is indeterminate. The point is that, taking the native's observational beliefs to be the set of sentences \( \{ O_H \} \) of our home language, under the determinate scheme of translation, we can construct \( T^1_H \) and \( T^2_H \), two different theories in Home both of which imply \( \{ O_H \} \), so we can translate the native's theoretical beliefs \( T_N \).
either as $T^1_H$ or as $T^2_H$, salva whatever constraints of usage, rationality and charity translation is required to observe. And moreover, thinks Quine, $T^1_H$ and $T^2_H$ make contrary ontological commitments, since their conjunction implies a contradiction.

Clearly, if the argument is to work, we cannot assume a warrant for preferring $T^1_H$ to $T^2_H$, or vice versa. So, since $T^1_H$ and $T^2_H$ are to be contraries, $T^1_H$ must imply some $P$ and $T^2_H$ imply $-P$, where $P$ is a theoretical sentence of unknown truth value. And this is what Quine had in mind when he wrote 23:

Are there two sentences [translations of a given native sentence] about which the speakers would disagree in truth value? No; the point is that the two translations of a native theoretical sentence may have unknown truth values, known only to be unlike.

The realism of Quine's presentation, the assumption that the two sentences have unknown truth values, is not necessary to his argument. It will be enough that the conjunction of $T^1_H$ and $T^2_H$ implies a contradiction to establish that they have different conceptual contents, and hence that translation, which is taken to preserve conceptual content, is indeterminate.

Arguably, an anti-realist construal of Quine's Duhemian argument implies a Popperian view of the epistemic relation between observation sentences and theory. Anti-realists suppose that the evidence for a disjunction $A \lor B$ must be either evidence for $A$ or evidence for $B$. So, an anti-realist will take $\{O_H^1\}$ to warrant $T^1_H \lor T^2_H$ only if either $\{O_H^1\}$
warrants \( T^1_H \) or \( \{O_H\} \) warrants \( T^2_H \). But \( T^1_H \models -T^2_H \) and \( T^2_H \models -T^1_H \). So if \( \{O_H\} \) warrants \( T^1_H \) it counts against \( T^2_H \), and similarly, if \( \{O_H\} \) warrants \( T^2_H \) it counts against \( T^1_H \). To the extent that the evidence warrants either disjunct it exactly diswarrants the other disjunct. Since we are talking about the same body of evidence in each case, it seems that evidence does not warrant the disjunction. Rather, the anti-realist should claim that observation and experiment may falsify a theory (and all other theories with the same observable consequences), but not confirm a theory.

The Duhemian argument need not imply a retreat from a holistic theory of meaning. Given that \( T^1_H \) and \( T^2_H \) differ in conceptual content and members of \( \{O_H\} \) have the same conceptual contents as their images in \( \{O_N\} \), assuming that the translation (and therefore conceptual content) of \( T^1_N \) is indeterminate as between \( T^1_H \) and \( T^2_H \), and given that the conceptual content of Home is determined (to the extent that it is determined) holistically, and likewise the conceptual content of Native; given all this, we will require that there is another set of sentences \( T^!_N \) of Native, which also imply \( \{O_N\} \), whose translation is also indeterminate as between \( T^1_H \) and \( T^2_H \), and such that the conjunction \( T^1_N \) and \( T^!_N \) implies a contradiction. Holism requires that, if the members of \( \{O_H\} \) and \( \{O_N\} \) have the same conceptual contents, then all the sentential inter-relations of the members of \( \{O_H\} \) in Home are isomorphic with those of the members of \( \{O_N\} \) in Native. Such isomorphism may occur under only one mapping between \( \{O_H\} \) and \( \{O_N\} \) coupled with either a mapping
between $\langle T^1_H, T^2_H \rangle$ and $\langle T^1_N, T^2_N \rangle$ or a mapping between $\langle T^1_H, T^2_H \rangle$ and $\langle T^1_N, T^2_N \rangle$.

Returning to the Duhemian root of the argument, $T^1_H$ and $T^2_H$ may be observationally equivalent in the actual world, but perhaps in another causally possible world they imply different sets of observation sentences? Now since we are holding the translation of observational vocabulary fixed we may find that $T^1_H$ implies $\neg \text{If } X \text{ then } Y$ whereas $T^2_H$ implies $\neg \text{If } X \text{ then } \neg Y$ where $X$ and $Y$ are composed entirely of observational vocabulary and the two conditionals are counterfactual. But if so we can easily determine which is the translation of $T^1_N$ by discovering whether the native assents to a translation of $\neg \text{If } X \text{ then } Y$ or to a translation of $\neg \text{If } X \text{ then } \neg Y$. So, it would seem, to retain the indeterminacy of $T^1_N$ as between $T^1_H$ and $T^2_H$, we require that every sentence constructed from the observation vocabulary, including counterfactual conditionals, which is a consequence of $T^1_H$ be also a consequence of $T^2_H$ and vice versa.

However, a complication now arises. We have $T^1_H$ and $T^2_H$ implying a contradiction, but also empirically equivalent in all possible worlds. But elsewhere Quine denies that observationally equivalent theories can fail to be logically equivalent. Let me first sketch Quine's argument, and then reject it.

He first constructs two observationally equivalent theories which are, prima facie, logical contraries. He imagines an "exhaustive encyclopedic formulation of
our total theory of the world", and then he imagines transposing the words "molecule" and "electron" in that theory, these being two terms taken from the theoretical side of the observation-theory divide. Then, since all the implicative connections between the observation sentences and the sentences containing the words "molecule" or "electron" in the old theory are matched by implicative connections with those words transposed in the new theory, old and new theories have the same set of observational consequences. Yet

the two theory formulations are logically incompatible, for the one attributes properties to molecules that the other denies of molecules and attributes to electrons.

However, in such a case, the logical incompatibility being resolved by the mere transposition of the terms in one of the formulations, we clearly have, Quine thinks, merely two formulations of the same theory.27

Quine then proceeds to cases where the two formulations have the same observation consequences, are prima facie logically incompatible, but we cannot find a transposition of the theoretical terms which will resolve that apparent incompatibility.28

Being incompatible, the two theory formulations that we are imagining must evaluate some sentence oppositely. Since they are nevertheless empirically equivalent, that sentence must contain terms that are short on observational criteria. But then we can just as well pick out one of those terms and treat it as if it were two independent words, one in the one theory formulation and another in the other. We can mark this by changing the spelling of the word in one of the two formulations.
Whence:

Both can be admitted thenceforward as true descriptions of one and the same world in different terms.

We should not, in such cases, assume that they are two formulations of the same theory, having the same conceptual content and making the same semantic claim about the world. To take a familiar example, if $T_H^1$ were Newton's laws plus "The sun is stationary in Absolute Space" and $T_H^2$ Newton's laws plus "The sun is moving uniformly in a straight line in Absolute Space", then $T_H^1$ and $T_H^2$ are prima facie logical contraries yet observationally equivalent. Quine could restore logical harmony by taking "Absolute Space" to be a different proper name in $T_H^1$ and in $T_H^2$. But then those names have different conceptual contents and $T_H^1$ and $T_H^2$ are committed to the existence of different entities. So $T_H^1$ and $T_H^2$ turn out to be different theories making different, but mutually compatible, ontological commitments.

However, I think we should reject Quine's attempt to show that observationally equivalent theories cannot really be logically incompatible. For it forces us into a holism that makes language unworkable. The sentences of a language do not face the tribunal of the world as an undifferentiated and unarticulated whole. For a language to be workable it needs articulating inference rules defined on the syntax of its sentences, be they formal rules like $P, Q \vdash P \& Q$ or 'meaning rules' like $\text{Red } x \vdash \text{Green } x$. Suppose these inference rules appear to validate a sequent $S_1, S_2 \vdash S_3$, where that validation rests in part upon some
predicate $P$ being a common syntactic component of $S_1$ and $S_2$. Given Quine's discussion above, the inference rules really do warrant that sequent only if the following is not the case:

$$\exists T_1 \exists T_2 (S_1 \epsilon T_1 \land S_2 \epsilon T_2 \land (\forall O)(T_1 \vdash O \leftrightarrow T_2 \vdash O))$$

For if the above is the case, the common predicate $P$ in $S_1$ and $S_2$ might not be two occurrences of the same syntactic unit, so far as the rules of inference are concerned. I.e., inference rules are not defined on the syntax of sentences per se, but on the syntax of sentences as parts of theories. And, of course, which other sentences form the rest of the relevant theories will depend upon the intellectual context.

Now consider our task of choosing between theories en route to building up a more comprehensive theory of the world. Suppose $T_3^3$ employs some of the theoretical predicates common to $T_1^1$ and $T_2^2$, but is a rival theory to the empirically equivalent $T_1^1$ and $T_2^2$, having different observational consequences. It turns out that we cannot take the prima facie syntactically defined inferential relations between $T_1^3$ and $T_1^1$, nor those between $T_2^2$ and $T_2^2$, as real inferential relations. For, depending upon how we resolve the apparent logical conflict between $T_1^1$ and $T_2^2$ - say by transposing terms in one, say by respelling terms in one as indicated by Quine above - there may be 'knock on' effects on the detailed inferential relations between $T_1^1$ and $T_3^3$ and/or between $T_2^2$ and $T_3^3$. Clearly, no syntactically defined inferential relations can be taken as more than prima facie, as we build our scientific
picture of the world.

Even supposing two theories $T^3_H$ and $T^4_H$ which are not observationally equivalent and appear to have clear syntactic inter-relations, those relations are provisional upon the later development of some $T^1_H$ and $T^2_H$, the discovery of their apparent logical incompatibility and observational equivalence, and consequent respellings and transpositions. In short, we do not know whether sentences which appear to be syntactically reticulated into the logical web of language really are so reticulated prior to the substantial achievement of our theoretical goals. This is what I meant by an unworkable holism.

Rejecting such a holism is consistent with a more modest holism. One can maintain that the articulating syntactic rules of inference are not to be justified piecemeal. The elimination and introduction rules for each logical constant need not be required to make a conservative extension to the atomic sentences and their consequences under the remainder of the rules. Rather, all that may be required of inference rules is that the whole system, the rules taken together, make a conservative extension to the atomic sentences. And the above is consistent with an empirical holism: Alternative distributions of truth values to the sentences articulated by the syntactic rules may correspond to the same empirical evidence, so the empirical import of a given assertion cannot be understood in isolation from the totality of sentences the subject would assert.
I hold no brief for such more modest holisms. My purpose is only to point out that rejecting Quine's unworkable holism is not to thereby reject all holisms.

Quine's proposal that observationally equivalent theories only appear to be logical contraries, amounts to a claim that such theories really belong to different languages (and all the observation sentences to which they are logically related count twice over, once as members of each language) whence they are not both in the domain of either syssem of syntactically defined articulating rules. Again, the effect of this proposal is that we do not know whether sentences really are syntactically reticulated into a net (= are members of the same language) prior to the substantial achievement of our theoretical goals.

Let us suppose that, pace Quine, we have two theories $T^1_H$ and $T^2_H$ which have the same set of observational consequences and which not merely appear to be but really are logical contraries. We return to the thesis that the translation of $T_N$ is indeterminate as between $T^1_H$ and $T^2_H$.

This thesis is plausible if we suppose that the indeterminate theoretical portion of $T_N$ consists of sentences whose primitive lexicon comprises theoretical terms only, plus co-ordinating definitions of the form $(x)(P_4 \leftrightarrow O_4)$, where $P_4$ is entirely theoretical and $O_4$ entirely observational, and $T^1_H$ and $T^2_H$ are similarly structured. But a glance at scientific theories shows that this supposition is unwarranted: the observational
vocabulary, whose translation is supposed to be determinate, penetrates beyond the co-ordinating definitions and into the heart of the theory. \( T^1_H \) may contain "The sun is stationary relative to Absolute Space" whereas \( T^2_H \) contains "The sun is moving uniformly in a straight line relative to Absolute Space", and in the determinately translatable observational portions of \( T^1_H \) and \( T^2_H \) we find, e.g. "The tower is stationary relative to the earth", "The stone is moving uniformly in a straight line relative to the tower."

The evidence that these are indeed the same relational expressions in both theoretical and observation sentences is that the syntactically defined inference rules of Home license parallel inferential relata centred on those expressions. And, of course, \( T^N_N \) will show a similar penetration of theory by observation vocabulary. But now we expect determinate translation: provided the logical contrariety between the theoretical statements of \( T^1_H \) and \( T^2_H \) involves some of the observational vocabulary, the determinate translation of the observation vocabulary of \( T^N_N \) into the common observation vocabulary of \( T^1_H \) and \( T^2_H \) will resolve the indeterminacy of the translation of the theoretical sentences of \( T^N_N \). The allegedly indeterminate theoretical sentences of \( T^H_H \) will contain some of the determinately translatable observation terms, and these will resolve the alleged indeterminacy.

For example, suppose \( T^1_H \) is the kinetic theory of gases, and \( T^2_H \) is a compressible fluid theory of gases, with identical observable consequences (PV=const) so far as either
theory has been developed. Whilst the theoretical portion of $T^1_H$ speaks of gas particles moving with greater velocity as it is heated, $T^2_H$ speaks of an increasing elastic force in a static fluid, where these are terms which recur in the observation portion of Home. So if the theoretical sentences of $T_N$ contain translations of the former observation terms and not the latter, then, ceteris paribus, it is a kinetic and not a fluid theory and $T^1_H$ is the correct translation.

The penetration of observation terms into theoretical sentences may be missed if we confuse the Duhemian argument with the indeterminacy of the set theoretical interpretation of natural number theory. Any $w$-sequence of sets will serve as an interpretation of number theory. But in this case the vocabulary of the arithmetic theory and the vocabulary of set theory are disjoint: not only are the numbers variously interpreted as one or other $w$-sequence, but the arithmetic operations on numbers are variously interpreted as operations on sets. Quine himself does not claim that the meaning of arithmetic theory is indeterminate as between its various set theoretical explications.\textsuperscript{30} Sure enough, some sentences about numbers take on different truth values under different set-theoretical explications of arithmetic - e.g. "($n)(n=$ the number of members of n)". However, thinks Quine\textsuperscript{31}:

the sentences about numbers which take on opposite truth values under different explications are sentences that have no clear truth values before explication.
Obviously Quine thinks the case is not one of indeterminacy of translation, but merely of *vagueness*: the language of arithmetic is vague relative to the language of set theory. But in contrast, there is no suggestion that the language of $T^1_H$ is any less precise than that of $T^1_H$ and $T^2_H$. Vagueness occurs when the inferential network of the one language is richer than the inferential network of the other. Indeterminacy when their inferential intra-relations are isomorphic, but are isomorphic under more than one non-equivalent mapping.

It is no mere parochial feature of 17c scientific theories that observational terms occur in remote theoretical sentences. In Chapter One I confessed for a prejudice for naturalism: psychological facts (including facts about the conceptual and semantic contents of sentences) are determined by but not reducible to physical facts. Wittgenstein persuasively argues that meaning must be manifest in use, manifest in our commerce with the observable features of the world. I conjecture that the *conceptual* contents of sentences are determined by those physical facts which are that commerce with observable features of the world. And that includes the *conceptual* contents of our *theoretical* terms - theoretical terms being, for the purposes of this discussion, terms for *unobservable* features of the world. So the conceptual content of theoretical language is somehow controlled, determined, by observable features of the world.
But a naturalistic attitude pulls in the opposite direction too: human beings are creatures with no special ontological status in the world, and with observational access to just certain features of it. Unobservable features of the world are not ontologically inferior to the observable features, and are, in the main, only contingently connected to observable features. Indeed, perhaps the observable properties are secondary properties - dispositional properties, objective relations between object and observer - and perhaps the theoretical primary properties are observer independent and only contingently related to observer-dependent secondary properties, since human observers are only a contingent feature of the world. So in our theorising we are hoping to describe observer-independent features which are only contingently connected to observable features. And yet our theoretical concepts are perforce controlled by our observational concepts.

It seems to me that our best hope for balancing out these opposing pulls lies in the deep penetration of theory by observational terms. If the conceptual content and semantic claims of talk of molecules being stationary or moving, or aggregating to form gross bodies, is the same as talk of billiard balls being stationary or moving, of bricks aggregating to form houses, then, if the theory is true, molecules are just as real as billiard balls and bricks, and the relationship between molecules and the gross bodies they form is just as contingent as the relation between bricks and the house they form.
We understand what is being claimed about molecules from our transactions with observable features of the world. But that understanding reveals those observable features as contingent, not merely in the Kantian noumenal/sense (the *bare* possibility of the world *not* being observable) but in the positive sense of opening up a whole range of other, observer-independent possibilities: ranging from an evenly diffuse gaseous world to a world in which all matter is aggregated into one lump.

Now, of course, the entities postulated by theory are strange in many ways, compared to observable entities. The difference between the observable and the unobservable is not merely a matter of scale: Had we Locke's microscopical eyes and a quick enough apprehension, the corpuscular world would not be the observable world scaled down and speeded up, and the corpuscularians realised so, with their doctrine of primary and secondary properties. But note our conceptual and ontological unease when the entities postulated by theory become too strange for any observation predicate to apply to them with literal meaning. The Ether is postulated as that which oscillates in the transmission of light. It is strange because massless and frictionless, but it becomes otiose and chimerical when the velocity of light appears the same to all observers no matter what their mutual uniform rectilinear motions. The lure of instrumentalism increases as the strangeness of the theoretical entities strains literal interpretations of the observation terms the theory applies to them.
Hence our unease about interpreting Quantum Theory realistically. How can the nature of an event (whether it is particulate or wavelike, but not both) depend upon another event when the theory forbids any causal route from the second event to the first, and there is no common causal determinant of both events?

I have argued that $T^1_H$ and $T^2_H$, though observationally equivalent, are logically and ontologically incompatible because of the different ways in which their common observation vocabulary penetrates into their theoretical sentences, and that we may expect to determine the correct translation of $T^1_N$ (given a fixed translation of the observation terms of $T^1_N$, and given that $T^1_N$ is observationally equivalent to $T^1_H$ and $T^2_H$) by examining how its observation terms penetrate its theoretical sentences. If $T^1_H$ and $T^2_H$ differ as a kinetic and a fluid theory of heat, for example, we expect no difficulty.

Putnam points to a more subtle way in which $T^1_H$ and $T^2_H$ may have contrary ontologies; differences which are not manifest in the penetration of theory by observation terms. $T^1_H$ and $T^2_H$ may differ because what one takes as a real theoretical entity the other takes as a logical construct, provided in each case what the theory takes as real is sufficient to furnish a coherent world.

Putnam gives a sketch towards an example: a one dimensional world in which creatures crawl to and fro along the line making local measurements to determine the large-scale geometry of their world. They come up with
three theories which differ in their ontologies. $T^1_H$ has an ontology of points, and segments, where "the same relation - 'part of' - holds between points and line segments which contain them, and between line segments and bigger line segments (and between any piece of the line and the whole line)." Note how the same relation 'part of' holds between observables and between unobservables. $T^2_H$ has an ontology of segments: points are logical constructs out of segments. "Point talk is highly derived talk about convergent sets of line segments." $T^3_H$ has an ontology of line segments with rational end points: "irrational line segments are treated as logical constructions - sets of 'points', where 'points' are themselves Cauchy convergent sets of rational line segments."

The point about this example, for my purposes, is that, in all three theories, the syntactical term 'part of' occurs in sentences ranging from observational right through to the most theoretical sentences about points. It is not obvious that an examination of the penetration of observation terms into theoretical sentences will distinguish one theory from the others. Hence, it seems we are left with the problem that the ontology and hence conceptual content of $T^1_N$ is indeterminate as between $T^1_{1-3}$. However, the relational expression 'part of' changes its conceptual content in $T^2_{1-3}$ (but not $T^1_{1-3}$) when we go from observational sentences to some theoretical sentences.
Consider the sentence "Point \( p \) is part of the segment \( s \)". In \( T^1_H \) this reports the physical part-whole relation between the real entities \( p \) and \( s \). In \( T^{2-3}_H \) points are abstract entities - certain sets - and in \( T^3_H \) \( s \) may also be an abstract entity, a certain set (if \( s \) is an irrational segment), or \( s \) is a real concrete entity. One could argue that a literal part-whole relation holds between one abstract entity and another - say, if the one is a member of the other, or a proper or improper subset of the other. But there is no literal part-whole relation between an abstract entity (a point in \( T^{2-3}_H \)) and a real concrete entity (a segment in \( T^2_H \), a rational segment in \( T^3_H \)). Hence, in \( T^{2-3}_H \) in some sentences of the form \( p \) is a part of \( s \), the term "part of" does not have its literal meaning, does not have the conceptual content that it has in observation sentences.

Conceptual content must be manifest in use. So we can ask what it is about \( T^{2-3}_H \) which manifests this change of meaning of "part of"? I.e. how do we, the speakers of Home, know of the disparate ontological commitments of \( T^{1-3}_H \)? Clearly because \( T^{1-3}_H \) differ precisely in their explicit ontological claims: \( T^1_H \) contains "Points, irrational and rational segments are all real parts of the world"; \( T^2_H \) contains "Rational and irrational segments are real parts of the world, but points are logical constructs", and points are explicitly or contextually defined as certain sets of segments (or sets of sets of segments, or whatever); and \( T^3_H \) contains explicit or contextual definitions of points
and of irrational segments as certain sets (of rational segments, or of sets of rational segments, or whatever). And now we have the further feature to look for in $T_N$, when trying to decide between the competing translations. We must look for explicit statements in $T_N$ about the different ontological statuses of points, rational and irrational segments, whether one is or is not contextually or explicitly defined in terms of the other. If there is no such, if nothing about $T_N$ manifests a change in sense of "part of" as it penetrates into the theoretical recesses, then $T_H$ is the correct translation.

Finally, let me mention that Putnam's purpose in constructing the three theories was different from mine: he was concerned to show that observation could not confirm one without confirming all, and hence that all three could be true (= ideally confirmed) together.
Appendix A: Empirical equivalence and ontological equivalence

Quine's way of resolving the apparent logical incompatibility between two observationally equivalent theories is surprising. More representative of the broadly empiricist tradition, in which I would place Quine, is the view that the logical incompatibility between $T^1_H$ (Newton's laws plus stationary sun) and $T^2_H$ (Newton's laws plus moving sun) is only apparent because a predicate or term essential to that apparent conflict ("Absolute Space") is empty of conceptual content and therefore without semantic content. Newton thought his term "Absolute Space" had conceptual content because he thought the only alternative 'coherent conception' of space was Leibnizian relativism in which all frames of reference were equivalent. (By a 'coherent conception' I mean, in this context, a conception which could have semantic content, a conception which describes a possible world. Newton worried about the coherence, in this sense, of his conception of gravitational action at a distance, and Boscovitch thought action at a distance the only coherent conception of action.) Newton had empirical evidence that there was more to Nature than the facts recognised by Leibnizian relations, because centripetal forces showed circular motions to be absolute. So the more, Newton thought, could only be Absolute Space, that being, he thought, the only conceptually coherent alternative.

However, later theorists found a conceptually coherent tertium quid: a certain class of reference frames are ontologically privileged, not all (pace Leibniz), but
none in that class was ontologically privileged over the rest (pace Newton). From the point of view of these neo-Newtonians, the talk of "Absolute Space" which makes $T^1_H$ and $T^2_H$ seemingly mutually incompatible is conceptually empty: the sentence "The sun is stationary in Absolute Space" in $T^1_H$ is merely an empty attempt to ascribe some additional ontological status to the sun-centred reference frame. The attempt is empty because we have no conception of what ontological extra is being bestowed on that reference frame by $T^1_H$. So when $T^2_H$ assigns some ontological extra, we know not what, to some other reference frame, we simply have no conception whether this is the same or a different unique peculiarity, and so no conception whether $T^1_H$ and $T^2_H$ are really in conceptual and therefore semantic conflict, or whether the dispute is merely verbal.

This way of showing that two empirically equivalent theories are not really logical contraries depends upon finding some third theory, which theory is (i) judged conceptually coherent (a picture of a possible world), and which is (ii) free of the putative ontological baggage which generated the apparent logical incompatibility between the first two theories. Famously, later generations have thought conceptually coherent, in my present sense, what their forefathers jibbed at.

However, we should not suppose that for any pair of observationally equivalent theories which are apparently logical contraries there is some third theory which will resolve their apparent logical conflict in this way.
The ontological "highest common denominator" may be too small to furnish a real world. As Putnam remarks of empirically equivalent but logically incompatible theories, and the attempt to factor out a conceptually coherent common denominator:

One can construe space-time points as objects, for example, or as properties. One can construe fields as objects, or do everything with particles acting at a distance (using retarded potentials). The fact is, so many properties of THE WORLD - starting with just the categorical ones, such as cardinality, particulars, or universals, etc. - turnout to be 'theory-relative' that THE WORLD ends up as a Kantian 'noumenal' world, a mere 'thing in itself'. If one cannot say how THE WORLD is theory-independently, then talk of all these theories as descriptions of 'the world' is empty.

Without going so far as Putnam, without supposing that the set of empirically equivalent but logically contrary theories have nothing in common, no common conceptual and ontological commitment, it may be that they do not have enough in common to furnish a world.

Putnam calls his response to this problem "internal realism". He identifies truth as "an idealisation of justification" and he explains:

the two key claims of such an idealisation theory of truth are: (i) that truth is independent of justification here and now, but not independent of all possibility of justification. To claim that a statement is true is to claim that it could be justified; (ii) that truth is expected to be stable, or 'convergent'; if either a statement or its negation could be justified, even if conditions were as ideal as one could hope to make them, there is no sense in thinking of the statement as having a truth value.

So it seems that both our logically incomptible but empirically equivalent theories are true, or neither are. And if they are both true, then they cannot really be
It seems to me that Putnam's internal realism amounts, in the end, to Quine's proposal considered in the chapter, and it suffers from the same unworkable holism: we have a system of articulating inference rules defined on the syntax of the sentences of the language, but given two sentence tokens whose syntactic forms seems grist to the inferential mill, we do not know whether the rules really apply, whether they are two tokens from the same language; that depends, not upon our intentions when producing them, but on how they stack up into large-scale theories: if they are members of two observationally equivalent but seemingly logically contrary theoretical wholes, then the syntactic rules do not apply to both together, they are not linguistic bedfellows. Only on completion of our search for the true theory, or theories, of the world do we know whether or not the same specific conception of truth applies to both tokens.

To reject internal realism is not to embrace "metaphysical realism", characterised by a verification transcendent conception of truth, even though Putnam presents them as jointly exhaustive alternatives. If \{T^i\} are a set of mutually contrary and jointly exhaustive (Though how could we ever know?) theories of the world, then rejecting internal realism is accepting that they really are logical contraries. But that does not commit me to one of them being unknowably true - metaphysical realism. All I am committed to is, supposing
\( T^i \vdash -T^j \), that if \( T^i \) is true, then \( T^j \) is false. I am not thereby committed to \( T^i \), say, being unverifiably true or being unverifiably false (i.e. to \( \vdash T^i v -T^i \)). Just so the intuitionist mathematician traces out the logical implications of Goldbach's conjecture, without thereby committing himself to the conjecture having a verification transcendent truth value.
Footnotes to Chapter Three

1 I take the classic arguments for indeterminacy to be Quine's "gavagai" argument in "Ontological Relativity" in Quine (3) and the "Duhemian" argument in Quine (4). Putnam's arguments I take from Putnam (4), Introduction vii-xii; Chapters 1 and 12. And Davidson's arguments I take from "Reality without Reference" and "The Inscrutability of Reference" in Davidson (3).

2 Notably the argument of Massy (1) and its discussion by Kirk (1) pp171-177, and Evans's demolition of Quine's "gavagai" argument in Evans (1).

3 This sentence is merely intended not to prejudge whether a conventionalist account of mathematical truth can permit Golbach's conjecture to have an unknowable truth value, and like issues.

4 Quine (4)
5 Putnam (3) p33
6 Putnam (4) p2, pp15-16
7 Putnam (3) pp33-35
8 Ibid. pp36-37
9 Putnam (4) pl8
10 Ibid. ppxi-xii
11 Putnam (3) p46
12 Ibid. p47
13 Putnam (4) pl7
14 Ibid. p24
15 Davidson (3) p223
16 Ibid p224
17 Davidson does not formulate the questions in these terms, but I take them to be implicit in his discussion
18 Ibid. pp235-237
19 Ibid. p235
20 I have in mind Burge (1) and Taylor (1)
21 Davidson (3) p75
22 Quine (4)
We do not even require that, were the condition $X$ fulfilled, there would be observers to determine whether or not $Y$. Maybe $X$ describes an observer-unfriendly world - e.g. a world in which there are no gross objects, in which all matter is diffuse. We only require that $X$ and $Y$ are constructed from logical constants and vocabulary taken from 'below' the agreed line dividing observation from theoretical sentences.

Quine (5) and Quine (6) Chapter Two

Quine (5) p320: two formulations express the same theory if they are empirically equivalent and there is a reconstrual of predicates that transforms one formulation into a logical equivalent of the other.

Quine (6) pp29 30

Occam's razor has been blunted: Quine has reconciled the seemingly rival theories by multiplying entities beyond empirical necessity.

In Davidson and Hintikka (1)

Davidson and Hintikka (1) p296

Of course, this does not commit me to the conceptual content of some theoretical term being some cluster of observable properties in the manner refuted by Kripke (2) and Putnam (1) Chapter 12, "The meaning of 'meaning'". But it does commit me to some Fregean view of conceptual contents - perhaps: that natural kind to which most of the colourless liquid, common to lakes and rivers, etc. belongs. Judicious handling of the scope of modal operators will enable us to say that $XYZ$, the colourless liquid, common to lakes and rivers, etc. on Twin Earth is not water. See Searle (2) Chapters 8 and 9.

Putnam (2) pp130-133

What follows is a risky hypothesis about Newton's psychology, based mainly on my understanding of the Clarke-Leibniz correspondence.
Field, in Field (1) and (2), proposes, as an example of indeterminacy that we can translate Newton's use of "mass" either as "rest mass" or as "relativistic mass". But proposed translations are wrong, I think. If we take Newton's use of "Absolute Space" to be empty, as suggested in the main text, then there may be no translation of Newton's use of "mass". On the other hand, if we expunge "Absolute Space" and adopt a neo-Newtonian theory in which there is a class of privileged reference frames, and if we take this as a coherent theory, one which describes a possible world, then it is simply an empirical fact that the actual world is Einsteinian, that objects have two properties, rest mass and relativistic mass, and not the property of neo-Newtonian mass. So, because the neo-Newtonian's mass is neither conceptually or semantically equivalent to either rest mass or relativistic mass, it should not be translated as either.

Compare: We are not tempted to translate "centaur" as either "man" or "horse".

Putnam (2) p133
Putnam (4) p84
Ibid. p85
Putnam, in Putnam (3) p73, says an internal realist would not object to two true theories being 'incompatible': "why should there not sometimes be equally coherent but incompatible conceptual schemes which fit our experimental beliefs equally well? If truth is not (unique) correspondence then the possibility of a certain pluralism is opened up." Internal realism, so understood, is incomprehensible to me: how can two true theories be incompatible (as distinct from logically independent or even incommensurable, because they contain different primitive terms)? So I have taken internal realism to imply that empirically equivalent theories are not (really) logical contraries, only appear to be incompatible.
Chapter Four: Narrow and Wider Functionalism

Section (i): Introduction

In this chapter I shall argue for wide functionalism by arguing against narrow functionalism. But first let us roughly distinguish narrow from wider functionalisms.

Naturalism requires that the physical facts constitute (many-one) and realise (one-one) psychological facts. Narrow functionalism is the view that the relevant physical facts are confined within the bodily envelope, plus sensory input and motor output across that bodily envelope. (Serious narrow functionalists confine their attention to the central nervous system, or some proper part thereof. Since I intend to reject narrow functionalism I can afford to set the boundary generously wide.) Wider functionalisms, on the other hand, take the constituting and realising physical facts to extend beyond the bodily envelope. How far beyond? I suggest, in Tom's case, to include all those physical entities which are the res of Tom's propositional attitudes. Remember, at the end of Chapter Two we pruned the objects of Tom's attitudes: although Tom referred to the car at the end of the assembly line, to Socrates, etc., it did not follow that those acts of reference manifested attitudes de that car, or de Socrates, etc. Rather, his attitudes were de the assembly line, and de the name "Socrates", for he intended to refer (successfully, as it happens) to whatever car was then at the end of the assembly line, to whoever is referred to by the name "Socrates" in that context. Generally, the
objects of our attitudes are those entities to which we have normal perceptual access and are the normal objects of our manipulative skills. — which includes our hearings of the name "Socrates", and our vocal productions of tokens of that name.

Section (ii): Fodor's narrow functionalism

Fodor remarks¹:-

so far as rational psychology [= narrow functionalism] per se is concerned, S [the subject] could be an ideal entity "with no links to the physical world". That is: whether the properties that the Rational psychologist attributes to a mental state are in fact properties of that state isn't allowed to depend essentially on the existence of such links.

However, in the subsequent discussion he seems to require, as a necessary condition of psychological states, nomic links between internal states and their external semantic contents. He requires that the play of internal states must preserve some designated semantic property, if that play is to qualify as a rational computation. Thus he writes²:-

Computations (unlike proofs) don't have to be truth-preserving — though Evans is right that there is something semantic that they have to preserve. Who would deny it?

The way to reconcile these two remarks is to say that the narrow functionalist presupposes that the play of internal representations do in fact preserve designated semantic values, but determining the conceptual contents of representational states and the rules of inference by which representational states beget representational states do not require him to determine their external links and semantic values.
We can now sharpen the distinction between narrow and wider functionalisms. Consider Tom-in-a-vat, where the input and output nerves are linked up to some machine. Let us suppose that, within Tom-in-the-vat (hereafter "TV") all that happens during the relevant period is molecule-for-molecule identical to what happens within Tom's bodily envelope. So, whatever narrow functional systems and histories those physical events constitute in Tom, they constitute also in TV. However, suppose that the activity of the machine which mediates between TV's output and input is caused by a troup of monkeys intent on monkey business. It just happens that they land on the levers, swing on the wires, in such a way as to mimic input relative to output such that what happens within TV's envelope replicates what happens within Tom's envelope.

In Tom's case there are nomic links between his internal representational states and their semantic values, which constitute a wide functional system. Had the 'start up' conditions been different (the objects in Tom's environment been somewhat differently disposed) the total wide physical set up would still have realised the same wide functional system - the same Turing Machine Table - though he would have enjoyed a different sequence of specific states of that system. But had the knobs on TV's machine been differently set, the monkeys would not have adjusted their dance accordingly. Thus TV is not a proper part of the same wide functional system as Tom.
So a wide functionalist will deny that TV enjoys the same mental life as Tom.

However, the narrow functionalist has no grounds for such discrimination. Tom and TV are narrow replicas, physically and therefore functionally. Both get the same inputs and cough up the same outputs at their respective envelopes. Certainly they differ in that Tom inhabits a world which is wont to provide inputs listed on his narrow machine table, whereas it was just luck that TV got inputs listed on his narrow machine table. But this difference does not prevent TV from also realising the narrow functional system. Compare a Turing Machine whose relevant input is a tape with boxes in which "1" or "+" have been ascribed. The question whether that machine is likely or unlikely to be provided with such inputs is irrelevant to the fact that it realises that machine table. No doubt, if baby scribbles on the tape, or worse, and then forces it through the machine, it will break, it will no longer realise that machine table. So no doubt the most likely thing to have happened to TV would have been a 'nervous breakdown'. But that didn't happen, because, as it happened, the monkey-play produced an input which matched output à la machine table. So the narrow functionalist has no grounds to deny TV the mental life of Tom.

In the light of this, I think we should understand Fodor's narrow functionalism as follows: The narrow functionalist presupposes that the internal representations
have semantic values which are preserved by the inference rules defined on the syntax of those representations. But this does not require any wide functional system or nomic relations external to the bodily envelope. Although TV's representational states do not represent his environment, he is hallucinating that he is a Tom, his inference rules preserve designated values in that they never lead from premisses with designated values to a conclusion whose value is not designated. And not only do they not, de facto, but they cannot, no matter how Tom is embedded in whatever possible world. Hence the narrow functionalist does not need to consider what lies beyond the bodily envelope; however it is out there, designated semantic properties are preserved by computations.

Fodor's main premiss for his narrow functionalism is his so-called "formality condition". Cognitive psychology aims to show that mental processes are computational processes on mental representations. (Or, more modestly, Cognitive Psychology aims to model human abilities by computational programs.) Fodor argues that inferential (i.e. computational) operations are formal in the sense of being defined on the syntax of mental representations.

And what is the syntax of a mental state? Fodor writes:

> What we want to develop is a substantive account of formality (a theory of what counts as a syntactic description of a mental representation; hence a theory of the descriptions to which mental operations have access) which has the consequence that syntactically identical mental states are functionally identical.

Fodor argues that because mental operations are defined on the syntax of mental representations (so understood),
they have "no access to" the semantic contents of those representations. Since the semantic contents are extra-craniial, narrow functionalism follows from the formality condition, thinks Fodor. The cognitive psychologist does not deny that his mental representations have extra-craniial semantic contents, he merely does not need to consider them in constructing his syntactic theories of computations on mental representations.

Fodor's formality condition does not imply narrow functionalism, and is quite consistent with wide functionalism. Fodor has conceded that the syntax of a mental representation is determined by its functional role: syntactically identical mental states are functionally identical. So, in claiming that the cognitive psychologist can determine the syntax of a mental representation without determining its semantic content, Fodor has assumed that the relevant functional system stops short of that semantic content. Thus Fodor's argument begs the question against wide functionalism.

Fodor is led to beg the question against wide functionalism because he construes mental representations on a sentential (rather than a relational) model, as a language of thought. However, we have already seen (Chapter Two, section (iv)) that a relational account of syntax is more plausible. A neuron may have the content \( P \lor Q \) because it is stimulated by a neuron with content \( P \) and by a neuron with content \( Q \). When that neuron fires in the course of a piece of reasoning of the pattern \(-P, P \lor Q \rightarrow Q\),
it still has the content PvQ, even though the causal routes P to PvQ and Q to PvQ are both dormant. And it has that content not because of its intrinsic structure, which need be no different from many another neuron with different syntactic contents, /because of its inter-connections with other neurons. Given such a picture of the syntax of that neuron, there is no bar, in principle, to the relevant relations stretching right out to the semantic contents of P and of Q - i.e. right out to stimulation by the P-ness of an object, and right out to stimulation by the Q-ness of an object. No matter that those receptors have that syntax even when stimulated by other causes, just as the neuron having the syntax PvQ has that syntax even when 'indirectly' activated by collateral information in the course of -P,PvQ -Q. Certain canonical routes of stimulation determine the syntax of the various neurons, and what makes these routes canonical may be that they start from their respective semantic contents. Thus, mental operations having access to syntax, does not preclude mental operations having access also to semantic contents.

Fodor himself provides us with a precedent! He argues elsewhere⁴ that high-level programming languages are not translated into machine languages by the process of compilation⁴:-

the referring expressions of ML typically name, for example, addresses in the machine, formulae stored or storable at addresses, etc. And the relational terms typically express operations that the machine can perform (e.g. writing down symbols, erasing symbols, moving symbols from addresses to addresses, etc.)
the sentence (capturable in ML) "The expression 'BOISE' appears at the address CITIES" is not remotely a translation of the sentence "Boise is a city" a sentence of the programming language. To suppose so is to commit a notably unsuitable version of the use/mention fallacy.

The point to note is that the semantic content of ML is confined within the envelope of the machine. How come the sentences of ML have the syntax they do? That syntax is determined by their roles in some functional system, and it turns out that this functional system includes their semantic contents! Fodor writes, with trivial change of example:

there is a consistent interpretation of ML under which the formula 'move the tape' is associated with the compliance condition moving the tape; and, moreover, it is a fact about the way the machine is engineered that it does indeed move the tape when it encounters that formula. This parallelism between the causal structure of the machine and the semantics of ML under its intended interpretation is what makes it possible to 'read' the machine's changes of physical state as computations.

It is because the relevant functional system which determines the syntax of the internal states of the machine includes their semantic contents, that computers are reliable. If the functional system in which the tokens of ML play their part stopped short of their semantic content, then the running of the ML program could not guarantee compliance with the ML program, where compliance is the semantic value to be preserved in the running of the ML program.

I shall argue that the functional system which realises the content of Tom's belief that Boise is a city is a wide
functional system, reaching out as far as the res of Tom's thoughts, though perhaps not to Boise, if he has never been there. A motivation for so arguing is that if a narrow functional system realised the contents of Tom's mental states, then that state would be as well indexed by "'Boise' is stored at the address CITY". So Tom would fail my criterion of Chapter Two for his mental states really having conceptual contents - they would be equally well indexed by two systems of indices with different conceptual contents. If so, so much the worse for my criterion, of course. But, even if we were to abandon that criterion, we would still have no explanation of why what it is like to be Tom is to have thoughts about Boise, cities, etc., rather than to have thoughts about what is stored where inside his skull. And I take it that it is an obvious datum that we do have thoughts about what goes on around us, and do not (usually) have thoughts about what goes on inside our heads. Our naturalism needs to incorporate that datum.

Section (iii): Narrow functionalism and the 'machine code'

My criterion for psychological realism was that Tom was of one mind, that there was something it was like to be Tom, cognitively and volitionally. Yet the same neural state of Tom can, it seems, have both the representational content that Boise is a city and the representational content that 'Boise' is stored at the address CITY. I hope that the neural state in question will constitute part of
a wide functional system and thereby have the representational content that Boise is a city, and also constitute part of a narrow functional system and thereby have the representational content that 'Boise' is stored at the address CITY.

I now conjecture that the relevant narrow functional system is not a proper part of the relevant wide functional system. The machine language computations of narrow cognitive psychology are not a proper part, confined within the bodily envelope, of the wide functional system which realises Tom's grasp and exercise of his conceptual repertoire. So the machine language concepts do not define the concepts which Tom grasps.

A simple Turing Machine will illustrate the relations I conjecture between wide and narrow functional states, between the machine code and the concepts which Tom grasps. Suppose the machine can add pairs of numbers, in the sense when presented, for example, with the tape where the arrow indicates the first box scanned, and A is its initial start-up state, it comes to rest with the tape as follows having gone through the following intermediate steps

1.  
2.  
3.  
4.  
The machine table, written in machine code, for our machine is:

<table>
<thead>
<tr>
<th>Tape entry</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IR A</td>
<td>IR B</td>
<td>bs C</td>
</tr>
<tr>
<td>+</td>
<td>IR B</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>blank space (bs)</td>
<td>?</td>
<td>bs LC</td>
<td>?</td>
</tr>
</tbody>
</table>

I've left blank those places which may be filled ad. lib., because they are not employed in such additions.

Only the initial state of the tape and the final state of the tape are syntactic representations of numbers, the intermediate states 1 to 8 do not represent arithmetic computational steps en route from 2+3 to 5. If there is a power failure and the machine stops at intermediate stage 4, say, this fails to represent anything arithmetic. It does not represent 6 as an intermediate conclusion in some wierd arithmetic computational route from 2+3 to 5.

From the point of view of the wide (arithmetic) functional system, only the initial stage and the final stage represent. It is a one stage arithmetic computation, and the intermediaries are just the levers effecting the transition.

From the point of view of the machine code, however, we may regard the entries in the boxes (1,+,bs) and the
internal machine states \((A, B, C)\) as sentences (an entry "+" says that "+" is stored at that address; the box being empty says that the box is empty; being in internal state A says that the machine is in internal state A, etc.), and we may regard the entries in the body of the table as inference rules - inference rules which not only dictate the immediate conclusion but also specify one of the premisses for the next step and say where the other premiss is to be found. And the semantic property preserved by these rules is compliance: the actual machine is so engineered that it complies with this table. From the point of view of the machine code, not only are the initial and the final state of the tape and machine significant, representational, but so are all the intermediate stages 1 to 8. All are steps in the machine code computation.

A peculiarity of this example is that those states of the total machine which are arithmetic representations (wide functional representations) are totally disjoint from those states of the total machine which are machine code representations (narrow functional states). For it is the whole tape at the initial stage, and the whole tape at the final stage that are arithmetic representations. Whereas the machine code representations are the content of the box being scanned and the current internal state \((A, B, \text{ or } C)\) of the machine at each of the stages. In other cases we expect wide and narrow functional states to partially overlap: the same state may represent that Boise
is a city in the wide functional system, and represent that "Boise" is stored at the address CITY in the narrow functional system.

This view of narrow functionalism and the machine code fits certain aspects of Fodor's formality condition: the rules of inference of the machine code in the body of the machine table are defined on the syntax of the sentences of the machine code (what is in the box being scanned, plus the internal state of the machine), and thereby, indirectly, on the syntax of arithmetic representations — for the task of the machine, what the machine code representations achieve, is the transformation of the description of a number in a form A+B into a numeral denoting the same number. But, we should note, that Fodor's conclusion that access to syntax precludes access to semantic content does not follow. True, the narrow functional system, the machine code computations, have no access, direct or indirect, to the arithmetic semantic content of certain total states. For it is we ourselves who declare (ex cathedra from the point of view of the Turing machine, so to speak) that the initial state denotes 2+3 and the final state 5, say. In other contexts they may denote other things. But, the machine code computations, by having access to the syntax of the sentences of the machine code thereby have access to their semantic contents.

Moreover, the machine code and computation does not determine which internal states in the machine computation
are 'wide' representational states. The first and last only, or some or all intermediate states also? The same states as are sentences in the machine code, or wholes made up of machine code sentences? And even if we do know which states of the machine code are 'wide' representations, this does not determine their 'application', does not determine whether they are arithmetic sentences or sentences about some other, quite disparate subject matter. So the concepts of the machine code and the machine code computations do not define the conceptual contents of the wide functional system. Judged by the analogy with our Turing machine, narrow functionalism does not look plausible. All that relevantly goes on within the thinker's head is capturable, directly or indirectly, by the program in 'machine code', and the concepts which the thinker grasps in thought cannot be defined in machine language. Perhaps those concepts are to be defined by a wider functional system centred on the thinker?

Section (iv): Narrow functionalism and rational inference

Functionalism seeks to assign conceptual contents to our thoughts in such a way that the unfolding sequence of thought is revealed as a rational process of inference; as computations, to use Fodor's term. Any functional system operates subject to certain boundary conditions, and the specification of those boundary conditions, the servicing and maintainance manual for the machine, is not part of the functional theory, because they vary from one
realisation of the theory to another. Within the functional theory the boundary conditions can only be stated tautologically: the machine must be such as to realise the functional theory. In the case of narrow functionalism, whether the creature realises the functional theory or not does not depend upon what happens outside the bodily envelope, but only upon what goes on inside. The creature's thought processes involve their conceptual contents and are rational, to whatever extent they are rational, however it is embedded in whatever world. To a wider functionalism, however, the creature's interaction with its environment may be a necessary constituting part of its rationality, a necessary constituting part of its logical conceptions.

I claim that rational thought presupposes wide functionalism. To support this claim I shall show that a narrow functionalist account of existential quantification provided by Field is defective. His narrow subjective semantics for classical logic cannot discriminate between reading "∃" as "there is" and reading it as "there may be", with variable modality. To disambiguate we need to examine what kind of world the creature inhabits and how it interacts with that world. Secondly, I shall consider work of Hintikka and Rantala. Again I shall show that whether a creature's quantifier rules are rational or not depends upon how it is embedded in what kind of world. In contrast to this, the narrow functionalist must suppose that the rationality of 'topic neutral' inference rules is independent of how the
creature is embedded in whatever possible world: they can never lead from true thoughts to false thoughts; the worst that can happen being that, due to stepmotherly Nature, massive hallucination, or whatever, they fail to lead from true thoughts to true thoughts.

Finally, in this section, turning to topic-specific, but nonetheless syntactically definable inference strategies, i.e., to local inductive rules, we shall see that the rationality of their employment is obviously dependent upon being favourably embedded in a user friendly world, upon a wide functional system and the maintainance of its boundary conditions.

First, then, Field and the existential quantifier. I shall show that Field's semantics for classical logic, given in terms of subjective probability and therefore acceptable to a narrow functionalist, fail to determine the sense of "∃". We could have two creatures, both realising the same train of thoughts bound together by the same inference rules as described by Field, but one means by "∃" that there is... and the other means that there may be..., depending upon their wider functional relations to their environments.

Field gives a semantics for a first-order predicate language with ¬, ∧, and ∃ primitive, based on a concept of subjective conditional probability, rather than on the usual notions of reference, extension and truth. McGinn

sees Field's work as providing the beginnings of a theory of sense as inferential role acceptable to narrow
functionalism. Field gives his own motivation for his paper and sums up its results thus\textsuperscript{12}:

Michael Dummett has suggested that unless one believes that sense can be made of the notion of truth, one should not accept the laws of classical logic; for the only rationale there could be for accepting those laws (as opposed, say, to the laws of intuitionistic logic) is that they are precisely the laws that preserve truth. In this paper I will establish a result which, I think, disproves this suggestion. The result is that, in addition to truth-theoretic semantics of the sort made popular by Tarski, there is a way of doing semantics in terms of a purely epistemic notion, viz. subjective probability; and that, with respect to such a probabilistic semantics, classical logic is both sound and complete.

I do not question Field's completeness and soundness result. But I shall show that these semantics do not fix the sense of "\(3\)". So I claim that the inference rules of classical logic do not fix the sense of "\(3\)". The sense of "\(3\)" is determined by how that logic is applied in the world.

First the rules, \(r(1)\) to \(r(3)\), by which Field defines a rational subjective probability function \(P\).

\begin{enumerate}
\item \(P\) satisfies axioms \(\text{l to \text{VII}}\):
   \begin{enumerate}
   \item \(P(A/B) \geq 0\)
   \item \(P(A/A) = 1\)
   \item \((\exists A)(\exists B)(P(A/B) \neq 1)\)
   \item If \(P(A/B) = 1\) and \(P(B/A) = 1\), then \((\forall C)(P(C/A) = P(C/B))\)
   \item \(P(A \land B/C) = P(A/B \land C) \cdot P(B/C)\)
   \item \(P(A \land B/C) \leq P(A/C)\)
   \item \(P(-B/A) = 1 - P(B/A)\), unless \((\forall C)(P(C/A) = 1)\)
   \end{enumerate}
\end{enumerate}

\(r(2)\) For any variable \(x\), for any formula \(F\) of \(L\) containing only \(x\) free, any sentence \(C\) of \(L\), and any singular terms \(t_1, \ldots, t_n\) of \(L\):

\[P((\exists x)Fx/C) \geq P(F(t_1) \lor \ldots \lor F(t_n)/C)\]

where "by \(F(t_1) \lor \ldots \lor F(t_n)\) I really mean of course \(\neg(\neg F(t_1) \land \ldots \land \neg F(t_n))\)."
Field notes that (r2) does not force, e.g.,
\[ P((\exists x)(Fx \land \neg Fx)/T) = 0 \]
where \( T \) is any tautology, since (r2) merely requires that
the value of \( P((\exists x)(Fx \land \neg Fx)/T) \) does not drop below a
certain value, and so would be satisfied by setting that
value to 1. This leads Field to his last rule (r3), which
employs the notion of an extension \( L_+ \) of \( L \) obtained by
adding further singular terms to \( L \), and an extension \( P_+ \) of
\( P \) from \( L \) to \( L_+ \). Finally, by a reasonable extension \( P_+ \) of \( P \)
to \( L_+ \) Field means (i) \( P(A/B) = P_+(A/B) \) for all \( A/B \) common
to both \( L \) and \( L_+ \), and (ii) that the extension \( P_+ \) satisfies
(r1) and (r2) applied to \( L_+ \).

\[ (r3) \]
For any variable \( x \), and for any formula \( Fx \) of \( L \)
containing only \( x \) free, for any sentence \( C \) of \( L \),
and any number \( q < P((\exists x)Fx/C) \), there is an
extension \( L_+ \) of \( L \) and a reasonable extension \( P_+ \)
of \( P \) to \( L_+ \) such that
\[ P_+(F(t_1)v...vF(t_n)/C) > q \]
for some terms \( t_1,...,t_n \) of \( L_+ \).

Field goes on to prove that classical logic is sound and
complete for \( L \) relative to (r1) to (r3).

To do so he defines the probability of a set of
sentences \( A_1,A_2,... \) conditional upon \( C \) as
\[ P(\{A_1,A_2,...\}/C) = \lim_{n \to \infty} P(A_1\land...\land A_n/C) \]
Then he writes:

We can now say that an inference from a set of \( T \)
sentences to a sentence \( B \) is legitimate under
interpretation \( P \) if \( (\forall C)(F(B/C) \geq P(T/C)) \), and that
such an inference is probabilistically valid if it
is legitimate under all interpretations.

He then shows that classical logic, understood as a
syntactically defined set of inference rules, is sound and
complete when interpreted as a logic of probabilistic
validity, provided that any \( P \) conforms to (r1) to (r3).
So far, it seems, so good for narrow functionalism: the rules (r1) to (r3) constrain P in total ignorance of the semantic contents of the sentences of L, and even the senses of the primitive predicates and names of L are not specified; but, it seems, the rules (r1) to (r3) do serve to fix the senses of the 'logical' words \( \forall, \land, \text{and} \exists 
\). However, I shall argue, the sense of "\( \exists x \)" is left ambiguous between the desired "there is an \( x \)" and an unwanted "there may be an \( x \)", where the sense of "may be" varies with the number of quantifiers contained within the scope of the quantifier in question.

The key point is that (r3) only requires that there be an extension \( L_+ \) of \( L \) and a reasonable extension \( P_+ \) of \( P \) to \( L_+ \), where, for a narrow functionalist this can amount to no more than the bare possibility of consistently (with one's other beliefs) adding the requisite singular terms to \( L \), with no requirement that those additional singular terms actually do have extra-cranial bearers. And Field himself gives just such an interpretation of (r3)\(^4\):

In thinking about (r3), it is important to realise that it does not say that an agent with a probability function \( P \) must be in a position to introduce the new names that (r3) requires; it says only that such names could be introduced without disturbing the original probability function or inducing violations of (r1) or (r2) in the expanded language.

Obviously, Field wants this weak reading of (r3) because he is seeking a semantics of classical logic.

This impression, that the extensions \( L_+ \) and \( P_+ \) which are quantified over in (r3) are constrained only by the syntactic consistency of the new sentences, a consistency
defined by (r1) and (r2) applied to any resulting $P_+$, is confirmed by the following passages, where Field is talking about the "conceptual role" of a name in relation to (r3):  

by quantifying over conditional probability functions defined on extensions of the language we're interested in, we get the effect of quantifying over possible conceptual roles.

And in a footnote to the above he comments further:

The idea actually requires a bit more sophistication. For it is plausible to hold that conceptual roles behave holistically: if certain possible conceptual roles are "occupied" in a given language ... this places certain constraints on what other conceptual roles can [N.B.] be occupied simultaneously in that language. If this holistic thesis is right, then what we want to quantify over in determining the conditional probabilities of quantified sentences of $L$ isn't all possible conceptual roles, but only the conceptual roles "compatible with" the conceptual roles already occupied in the language.

[Field's emphasis]

And the "compatibility" in question of "Fa", where "a" is some new name, is going to turn out to be consistency according to the rules of classical logic with the rest of one's beliefs.

These remarks strongly suggest a non-standard interpretation of "[]" for a $P$ which conforms to (r1) to (r3); an interpretation in terms of the mere possibility of an object, rather than the actual existence of an object.

For ease of exposition let us introduce universal quantifiers into the language via the usual classical definition: $(x)Fx = \neg(\exists x)\neg Fx$. We can then regard every sentence of $L$ as having the same sense as its classically equivalent sentence in normal form with all quantifiers gathered into a prefix. Suppose we are then given a $P^0$ defined on the closed quantifier free sentences $L^0$ of $L$. 
We can extend that $P^0$ to $P^1$ defined on $L^1$, which is $L^0$ plus all the closed sentences of the forms $(\exists x)Fx$ and $(x)Fx$ with **one** initial quantifier. And we can require of, e.g., $P^1((\exists x)A/B)$ only that it has the **maximum value** of any $P^0(A(t_1)v...vA(t_n)/B)$, where $t_1,...,t_n$ are old or new names, **consistent with** $P^0$ and (rl) to (r3). E.g., we are adding $(\exists x)A$ to our stock of beliefs if and only if some $A(t_1)v...vA(t_n)$, where $t_1,...,t_n$ are old or new names, can be **consistently added** to our former stock of beliefs, where consistency is defined, in effect, by the rules of classical logic. In this way we are reading $(\exists x)A$ and $(x)A$ as, respectively: $\Box_0(\exists x)A$ and $\neg\Box_0(\exists x)\neg A$.  

Now, given $P^1$ defined on $L^1$, we can extend $L^1$ to $L^2$ by adding all closed sentences of $L$ with **two** initial quantifiers. And we extend $P^1$ to $P^2$ so as to read, e.g., $(\exists x)(\exists y)A$ as $\Box_1(\exists x)\Box_0(\exists y)A$, where $\Box_1(\exists x)$ indicates that, for some old or new names $t_1,...,t_n$, some $\Box_0(\exists y)A(t_1)v...v\Box_0(\exists y)A(t_n)$ is **consistent with** the conditional belief system $P^1$ defined on $L^1$. In this way we can recursively define a valuation of $P$ on $L$ such that, for example $(x)(\exists y)(z)(\exists w)Fxyzw$ takes the sense of $\neg\neg\Box_3(\exists x)\neg\Box_2(\exists y)\neg\Box_1(\exists z)\neg\Box_0(\exists w)Fxyzw$ where each original quantifier has a modal interpretation, and the modality varies with the 'depth' of the quantifier. Details of such a valuation are given in Appendix A to this chapter.

Given a narrow functional system whose inter-relations
realise the syntax of a first-order language, because that syntax is associated with some P that accords to (rl) to (r3), can we tell what the sense of "a" is, whether it is "there is", or "there may be", with variable modality?

Suppose Tom does not add a belief that ($\exists^i x)Fx$ to his stock of beliefs whenever he computes that a belief of the form $Fa$, where "a" is some new name, is consistent with all his beliefs of order i-1. (We measure the order of a belief by the number of nested quantifiers in its expression.) Then we have reason to suppose that Tom does not understand his belief indexed by ($\exists^i x)Fx$ as the belief that there may be an x such that Fx, in the weakest most permissive sense of "may", a sense which amounts to mere syntactic consistency with the rest of his beliefs of order i-1.

But looking inside Tom's head will not tell us whether he understands "$\exists x" as "there is" or as "there may be" in some stricter, more demanding modality: say 'real' possibility rather than mere logical possibility. To tell the difference, I suggest, we will need to see which interpretation of the syntax makes him the more rational agent in relation to his environment. I see no way of determining by looking inside Tom's head whether his judgements that ($\exists x)Fx$ are rash (read as: There is an F), or overcautious (read as: There may be an F).

Suppose we find that, relative to his environment, Tom is an efficient 'seeking and finding machine' for things of kind F. And suppose that, where ($\exists x)Fx$ is not
a theorem of classical logic, then Tom does not form the belief that \((\exists x)Fx\) unless he has an effective strategy for finding a substitution instance. That would convince us to read "\(\exists\)" as "there is". And that would require looking at the wide functional interaction between Tom and his environment.

Of course, this last possibility raises a host of knotty problems. If Tom is to believe that \((\exists x)Fx\), but without having any singular belief of the form \(Fa\), then to have an effective strategy is not to effect that strategy. So does Tom need to have effected an epistemic strategy which has shown him that he has an effective strategy? And what might strategies to disclose strategies be like? Also, if the detective is to believe that there was a murderer, yet the crime remains stubbornly unsolved, must he be in a position to know that someone has an effective strategy (perhaps the murderer!), or can the detective’s effective strategies embrace counterfactual possibilities?

I do not propose to pursue these questions. My purpose is only to show that they are questions about our normal epistemic powers, and as such they will raise the question of what kind of world we live in and what is our functional relation to that world. Given answers to such questions such that we can say that Tom does understand "there is an F" by his belief that \((\exists x)Fx\), we will be able to find a within-the-bodily-envelope replicate, Twin Tom, whose world is less user-friendly so that the
same strategies are only effective to support a "there may be an F". And of Tom-in-the-vat we will have no means of determining what he understands by a belief that \((\exists x)Fx\).

Work of Hintikka\(^9\) and Rantala\(^10\) shows that, given a non-standard model-theoretic interpretation, the inference rules of classical logic fail, leading from true premisses to false conclusions. In itself this is not surprising, for the models in question are not standard. But Hintikka claims that the senses of the logical terms are the same in these non-standard models as in the standard models. I shall argue that he is wrong on this point: the meaning of the quantifier changes from standard to non-standard models. However, suppose Tom employs classical inference rules, but is so related to the world that his thoughts have the non-standard semantic interpretation. Then Tom's inferential practices will be invalid and Tom is irrational. We cannot say that, from Tom's point of view he reasons rationally, because he assumes that he is so related to the world as to give his thoughts standard semantic interpretations, so he is merely labouring under a false assumption about what his world is like. He takes his inference rules to be necessary and not subject to contingent assumptions. There need be no room in his web of belief for the real relation between his inferences and his world. Tom is not merely wrong like someone who has made a factual mistake, but is nonetheless rational for that. Tom is a radical semantic misfit.

A wider functionalist, on the other hand, has the
means to rule out such radical semantic misfits. The relevant functional system which determines the inference rules, be they classical or of some restricted logic, include the subject's interaction with his world. So if the organism-world relations realise a non-standard semantic interpretation, then they determine non-classical inference rules also, and rationality reigns.

Hintikka starts from the folk-psychological common-place that we often fail to believe the less obvious syntactic consequences of our beliefs, and even, on occasion, positively hold syntactically contradictory beliefs. Supposing that a person believes the thought (sense) expressed by the obliquely occurring sentence in a correct belief report, then the problem, as Hintikka sees it, is that \( p \rightarrow \neg q \) would show that "Tom believes that \( p \)" and "Tom believes that \( \neg q \)" cannot correctly report the ideoelectic senses of Tom's beliefs. (Because, in more detail, Hintikka takes the oblique clauses to determine the set of Tom's epistemically possible worlds, and syntactic contradictions would together determine that Tom had the null set of epistemic possible worlds. Whence his behaviour becomes inexplicable.)

Hintikka suggests a solution. We need to distinguish between some proper part of the functional system which canonically determines the senses of the logical constants, and the remainder of the functional system which may involve the misuse of those logical concepts. Then syntactically inconsistent beliefs are shown to be
possible, and to be semantically inconsistent, as common sense supposed.

Given a first-order predicate language, Hintikka divides sentential implications into two mutually exclusive kinds: trivial inferences and non-trivial inferences. Hintikka holds that trivial inferences depend upon the senses of the premisses and conclusion, in a way in which non-trivial inferences do not, and that competent language users cannot make mistakes about trivial inferences, whereas they can about non-trivial inferences.

To distinguish trivial from non-trivial inferences Hintikka defines the depth of a sentence: roughly it is the largest number of nested quantifiers in the sentence. (Roughly, because one will need to ensure that the nesting is essential to the truth conditions of the sentence, and because the presence of proper names may increase the depth of a sentence.) He then defines a trivial inference as one in which no intermediate line has greater depth than that of any premiss or that of the conclusion. Whereas a non-trivial proof goes via an intermediate line of greater depth than any of the premisses or the conclusion.

Taking $p \vdash q$ to be a trivial inference of depth $d$, Hintikka suggests:

the following procedure for spelling out the meaning of $[p]$ ... we transform $[p]$ into its distributive normal form of depth $d$. The same goes for $[q]$. Grasping the meaning of $[p]$ and $[q]$ ... is closely related to being able to do this (in principle). But if our man can do this, and if $[q]$ is indeed trivially implied by $[p]$, then he cannot (in principle)
fail to see that the alternatives admitted by \([p]\) are all among those admitted by \([q]\), too, i.e. that \([q]\) is implied by \([p]\). The only reason he does not do this is his failure to appreciate fully what he says he believes and disbelieves.

Precisely what Hintikka means by "distributive normal form of depth \(d\)" need not detain us because two points are surely clear: firstly, the passage quoted does not warrant Hintikka's claim that competent language users cannot be mistaken about trivial implications. For grasping the senses of \(p\) and \(q\) only requires that the believer be able to carry out a certain canonical procedure which would reveal the implication, not that he has done so. So it seems that he can believe that \(p\) and fail to believe that \(q\), and grasp the meanings of \(p\) and \(q\), provided that if he were to carry out the canonical procedure he would discover his error. In fact, secondly, the way Hintikka connects trivial implications to sense does, at least prima facie, solve the functionist's problem. For to grasp the senses of \(p\) and \(q\) is to be able to carry out a certain canonical procedure competently (reduction to distributive normal form). This requires that an appropriate mechanism be realised in the functional system, but not that it actually be activated in any given case. So a subject can functionally realise the belief that \(p\), and fail to realise the belief that \(q\), provided he would discover his error if the mechanisms were activated. This is an account which makes trivial error functionally possible.

However, let us turn now to Rantala's non-standard semantic interpretation and the resulting failure of
non-trivial inferences. For it is from this that the
difference between narrow and wider functionalisms will
emerge. Rantala asks us to suppose that the domain of the
quantifiers are balls in an urn: to test a given sentence
for its truth value, we make successive draws (with
replacement) of balls from the urn, and take these as the
values of the variables in the sentence. In more detail,
given a sentence $S$, Myself (the truth proponent) and
Nature (the falsity proponent) play out a semantic game
on $S$, choosing sub-formulas of $S$ to eliminate the
constant ("$\&$", "$v$", or "$-$") of widest scope, and choosing
a ball from the urn to eliminate a quantifier of widest
scope, until we are left with an atomic formula and a
sequence to satisfy it (Myself wins) or not (Nature wins).
Truth of $S$ is defined as Myself having a winning strategy
on $S$, and falsity as Nature having a winning strategy.$^{21}$

In standard urn models we take the domain as fixed,
but Rantala introduces the idea of a $d$-invariant urn model:
such a model includes a mechanism which is triggered after
the players have between them made $d$ draws from the urn,
the $d^{th}$ draw triggers this mechanism which changes the
population of the urn, either completely or partially, by
adding and/or subtracting balls, before the $(d+1)^{th}$ draw.
Rantala allows the players to know of this mechanism,
and to know what change in the population its operation
will produce. Hence players may plan their game-strategies
with the change of population in mind. (Rantala's games
are games of perfect information.)

Rantala and Hintikka prove that, where $p$ trivially
implies \( q \), and \( p \models q \) is of depth \( d \), then \( p \models q \) holds in every \( d \)-invariant urn model. But where \( p \) non-trivially implies \( q \), and \( p \Rightarrow q \) is of depth \( d \), then \( p \models q \) fails in some \( d \)-invariant urn model. Clearly, what happens is that where \( p \) non-trivially implies \( q \) the proof contains a line of depth greater than \( d \), and this line turns out to be false in the \( d \)-invariant urn model because of the change in population triggered by the \( d \)th draw. The failure of this line permits \( p \) to be true and \( q \) to be false in that \( d \)-invariant urn model.

Hintikka wishes to use Rantala's non-standard urn models to "make model theoretic sense of" our failure to believe the non-trivial consequences of our beliefs. As a presupposition of this project, Hintikka claims:

Of course these logical consequences must be based on the classical (standard) interpretations of connectives and quantifiers. Thus an attempted non-standard interpretation is either bound to be beside the point or else to destroy the problem instead of solving it.

And, of Rantala's non-standard urn models, Hintikka claims:

Rantala's game-theoretical truth-definition for urn models is to all practical purposes the same as the game-theoretical truth-characterisation for old (invariant) models. No non-standard interpretation of logical concepts is involved here.

This step is crucial to Hintikka's argument because it allows the trivial inferences to determine the senses of the logical constants, senses which are not affected by the failure of non-trivial inferences in \( d \)-invariant urn models.

This last claim of Hintikka's is mistaken: the senses of the quantifiers, as specified by their game rules, are
different in games played with non-standard urn models from games played with standard urn models. The game rules for quantifiers in games with standard urns are:-

(1) (GU/GE) If $S$ is of the form $(x)S_0/(\exists x)S_0$, then Nature/Myself chooses a ball from the urn, and the game is continued with respect to $S_0(n/x)$, where $n$ was the ball chosen.

Now we could rewrite this rule, without change of force, as follows:-

(2) (GU/GE) If $S$ is of the form $(x)S_0/(\exists x)S_0$, then Nature/Myself nominates a ball which was, is, or will be in the urn, and the game is continued with respect to $S_0(n/x)$, where $n$ is the ball nominated.

Nominating a ball will do as well as extracting and replacing it, and since, for standard urn models, the population is invariant, balls which Nature/Myself can now nominate are precisely those which were, are or will be in the urn (for the duration of the game).

Rantala's quantification rules for games with his non-standard models are, slightly simplified as follows:

(3) (GU/GE) If $S$ is of the form $(x)S_0/(\exists x)S_0$, and $a_1, \ldots, a_i$ have been drawn earlier, then Nature/Myself draws an $a$ which is obtainable relative to $a_1, \ldots, a_i$, and the game is continued with respect to $S_0(a/x)$.

In effect Rantala's rule is:

(4) (GU/GE) If $S$ is of the form $(x)S_0/(\exists x)S_0$, then Nature/Myself nominates a ball which is presently in the urn, and the game continues with respect to $S_0(n/x)$, where $n$ is the ball nominated.

For what balls are obtainable relative to $a_1, \ldots, a_i$ drawn
earlier, are precisely those which are in the urn after
$a_1, \ldots, a_n$ have been drawn and replaced, which are precisely
those balls in the urn at that particular stage of the play.

It is obvious that (2) and (4) are different rules, and hence their respective equivalents (1) and (3) must be
different rules, bestowing different senses on the
quantifiers, because, if we take Rantala's $d$-invariant urn
models, but play the game according to rule (2), then all
non-trivial inferences are valid, whereas if we play the
game according to rule (4) they are, as Rantala shows,
invalid. So they cannot be merely two formulations of the
same rule.

Hintikka was concerned to give a semantic account of
inconsistent beliefs, and I explore the question of what a
functionalist should make of syntactically inconsistent
beliefs in Appendix B to this chapter. My purpose now is
to show the problem which Hintikka and Rantala's work generates
for narrow functionalism. The narrow functionalist, as
Fodor remarked, presupposes that the appropriate semantic
property of his subject's representational states is
preserved in computations, that the subject's lucubrations
are rational, that the syntactically defined inference rules
never go from true premisses to false conclusions - though
the subject may, perhaps, on occasion fail to apply, or misapply those inference rules so as to hold syntactically
inconsistent beliefs, and though in the limiting case of
the brain in the vat, the total hallucinator, truth is
preserved in the null sense that the premisses from which
false conclusions are drawn are themselves not true.

The problem for the narrow functionalist emerges when we compare Tom, a competent classical logician playing truth games with standard urn models, and his internal replicate Twin Tom who is playing with d-invariant non-standard models. Since they are replicas, we suppose that whatever games Tom wins/loses, Twin Tom wins/loses the corresponding games. But truth is having a winning strategy, not merely winning the game. There will be occasions where Tom is justifiably confident, having used his classical rules of inference to devise a winning strategy. And Twin Tom will be likewise confident, and does as it happens win that game by following that strategy, but he was just lucky. The game required more than d draws, his strategy could not guarantee a win because it did not take into account the possibility of the population changing, but his luck held; if there was a change it didn't show up in the subsequent draws. Thus Twin Tom is led by his classical inference rules, in cases of non-trivial inferences of depth greater than d, from true premisses to false conclusions, to the thought that if he has a winning strategy on these sentences then he also has a winning strategy on that sentence.

The wider functionalist has no such problem. Tom and Twin Tom realise different wide functional systems, are not psychological replicas because they are interacting with different kinds of urn. Tom's inference rules are classical, and a boundary condition for a wide physical system
realising classical inference rules is that the urn with which it interacts is standard, or that the subject nominates balls which were, are, or will be in the urn. Which is to say: the relations which determine the syntax of Tom's thoughts must reach out to entities which are such as to make the rules of inference defined on that syntax valid.

Twin Tom has access only to the then current contents of the urn, that urn being non-standard. Such relations, it seems, do not generate a classical syntax for Twin Tom's thoughts, just because the classical inference rules defined on that syntax would be invalid. But what, then, is it like to be Twin Tom, according to the wide functionalist, if it is not like being Tom? Allow me to postpone an answer until after we have considered the fallibility of inductive strategies. For the ability of inductive strategies to preserve designated semantic values also depends upon how the subject is embedded in what kind of world.

Fodor agrees that inductive strategies fall within the remit of cognitive psychology. For one of the generalisations which his narrow functionalist is called upon to explain is: -26

Beliefs whose content is that all x's are F are often caused by beliefs whose content is that some or other x is F

And he continues:-

There are many other such generalisations: logic can be viewed as (inter alia) formulating a number of them..., and so can decision theory, linguistics, and, in fact, every other area of cognitive psychology[N.B.] that has looked remotely like working
and yet: \(^2\)

Computations (unlike proofs) don't have to be truth-preserving - though Evans is right that there is something semantic that they have to preserve. Who would deny it?

The problem for narrow functionalism is that inductive strategies have to preserve the designated semantic value, yet whether an inductive strategy does so or not is a parochial matter, depending upon how the subject is embedded in what kind of world. I.e., it requires an examination of the links between a thinker and his world, a subject matter outwith the narrow functionalist's remit.

The designated semantic property for inductive inferences is, arguably, not mere truth, but something stronger: counterfactual reliability, perhaps truth across co-causal possible worlds. But whether a generalisation so learnt is a law or a mere accidental generalisation, whether it can support counterfactual conditionals, is a matter of the kind of world in which the creature is embedded. Tom may have the same counterfactual confidence in his belief that all x's are F's as his narrow replica does in another possible world, according to the narrow functionalist. But Tom's belief is nomically correct of his user-friendly world, whereas his replica's belief is nomically incorrect, it was just a coincidence in his possible world.

For example, suppose Tom is a cat, the pampered pet of an autocratic and wealthy matriarch. Then Tom may form the generalisation, based on particular satisfying encounters,
that humans are nice to him. Given the boundary conditions of Tom's world his belief is rational: visitors can be relied upon to vie for his favours. Twin Tom, on the other hand, although a narrow functional replica of Tom, is the pampered pet of a penniless inmate of a home for the elderly. By luck this change of circumstance has, in fact, had no deleterious effect on Twin Tom, he being a narrow replica of Tom. The other inmates are, as it happens, kind to cats. But Twin Tom's belief that humans are good to him lacks the semantic virtue of Tom's parallel belief: it lacks its reliability; it lacks any law-like status. The supervising authorities did not select inmates with a penchant for cats, it just happened that way. And Tom and his twin's beliefs about humans extends beyond present company - they would confidently approach any human they encountered.

Of course, Tom's generalisation does not hold in all co-causal possible worlds, as he would discover if his mistress die on him. But our inductive strategies often seek to preserve weaker semantic properties than causal necessity. However, when generalising from instances we do seek to preserve a semantic property stronger than mere truth: someone who inferred from some particular x's being F's to all x's being F's, but denied that there was any generative mechanism or counterfactual reliability, would have no warrant for his generalisation. And whether such inferences do have these semantic virtues depends upon
what the world is like in which the creature is embedded. Tom's inference patterns do preserve an appropriate semantic virtue, Twin Tom's do not. Tom's inductive strategy is a manifestation of his rationality, Twin Tom's is not.

Faced with our Twin Tom's should the narrow functionalist drop his requirement that computations preserve designated semantic values? Perhaps the narrow functionalist should say: We recognise other psychological subjects by recognising behaviour which is rationally attuned to its external environment; i.e., by recognising behaviour (human and animal) which manifests truth—or whatever other semantic value is designated—preserving computations. But, the narrow functionalist should say, that the computations are thus truth (etc.) preserving, is not a necessary condition of being a psychological subject, since truth (etc.) preservation depends upon how the creature is embedded in what possible world, matters beyond the remit of narrow functionalism.

Perhaps the minimal semantic condition should be: computations must be such that there is a possible world, and there is an embedding of the creature in that world, such that its computations are truth (etc.) preserving. But, likely, this is tantamount only to the creature's beliefs being syntactically consistent, and hence having a model.

Syntactic consistency alone is not enough, I think. The narrow functionalist requires the computations to be
truth (etc.) preserving because he wishes to describe the relevant functional system, determine its relevant syntax and the inference rules defined on that syntax; relevant, that is, to explaining the adaptive and computational success of our behaviour, to the extent that our behaviour is successful. The first-order physical facts confined within Tom's bodily envelope, or within some more restricted neural envelope, plus first-order in-and-output across that envelope, will realise many functional systems, whose inter-related states could be sententially indexed and their inter-relations described by inference rules defined on the syntax of those sentences. (A computer, being a universal machine, can model any functional system, and there will be a program whose sentences index the relevant states of the computer and whose 'inference rules' describe their functional inter-relations.) If we lift the constraint that the functional states have external semantic contents and the computations are (to the extent that they are) truth etc. preserving, we have no idea whether our computational psychology is modelling the relevant functional system.

The narrow functionalist assumes that he can identify the relevant functional system, and assign syntax and inference rules internally - i.e., in official total ignorance of the relation between those functional states and their semantic values whereby the computations are truth (etc.) preserving. Perhaps he believes this because he believes they are truth (etc.) preserving no matter how
the creature is embedded in whatever possible world. Or, more likely, because they believe the rules are truth (etc.) preserving in normal embeddings in this world - a normality he does not seek to investigate but presupposes. The narrow functionalist assumes that he can assign the appropriate syntax and define the appropriate inference rules on that syntax, by examining functional relations out as far as the bodily envelope, their being no need to follow those relations out to the semantic values of the internal states.

The plaint of this section has been that this is, more likely, not so. The meaning of the existential quantifier, the validity of classical inference rules and inductive strategies depend upon how the subject is contingently related to the semantic values of his thoughts. The relations which warrant the concept of existential quantification to certain states, and the relations which warrant assigning the syntax and inference rules of classical logic, or of an inductive strategy, those very warranting contingent relations stretch out to their semantic values, in normal successful cases. Only by linking up the internal states to their semantic values in normal cases, can the psychologist show that the internal states which he choses, the syntax and inference rules he assigns to them, are the relevant ones - relevant to explaining the creature's successful adaptation, learnt and innate, to its environment. And, of course, to include these relations in the overall functional system is to opt for a wider functionalism.
The wider functionalist need not deny that our Twin Toms enjoy psychological histories which are subjectively indistinguishable from their respective Tom's. Nor that they are as subjectively rational. They are like machines for grading apples that have been fed pears, but, as it happens, those particular pears were apple-shaped and were thus graded as apples. Twin Tom is not irrational in the sense of being at fault for thinking humans are disposed to be nice to him, and Tom Cat does not take credit for his generalisation having a certain nomic force, for Tom Cat has no knowledge of the influence of his patroness over other humans. Similarly, Twin Tom is not culpably ignorant relative to Tom of the kind of urn he is selecting from: For whatever Tom did or did not do to check of what kind was his urn, Twin Tom did also, and with the same result. But, the wider functionalist claims, to assign the syntax and inference rules we have to each Tom and his respective twin, is to take the Tom's relation to the world as normal (the standard) and his twin's as deviant. A distinction not available to the narrow functionalist.

The twins are in breach of the boundary conditions which determine whether or not the wide functional system applies by ensuring that inference rules preserve designated semantic values. So the twins are not in the same psychological states as their respective Toms. But we say that they are as a courtesy title because it is not evident to each twin nor to his Tom whether or not the relevant boundary condition is met.
Appendix A to chapter four

Field's (rl) to (r3) permit reading "∃x" as "there may be an x", where the relevant modality varies with the number of other quantifiers occurring within its scope. A reading of "∃x" is permitted, if values for P can be found which (1) conform to (rl) to (r3), and (2) would be plausible values, given that understanding of "∃x", where by "plausible" I mean only that the reading of "∃x" would agree with the values P assigns. So I propose to define recursively such a plausible P.

Suppose we are given P^0 defined on L^0, the quantifier-free closed sentences of L. It will be convenient to add "(x)" to the language L, and rules (r4) and (r5) to deal with it, so that we can regard each sentence of L as synonymous with its normal form in which all the quantifiers are gathered into a prefix. This will facilitate ordering the sentences of L into L^0, L^1, L^2, ..., L^i, ..., where L^i contains all the closed sentences of L with i or less quantifiers in their prefixes. We shall then extend P^0 to P^1, to P^2, ..., to P^i, ... recursively, in accord with (rl) to (r5), and such that, e.g.

(∃x)(y)(∃z)(w)Fxyzw

is plausibly read as

\( \Diamond_3(∃x) \rightarrow ∆_2(∃y) \rightarrow ∆_1(∃z) \rightarrow ∆_0(∃w) \rightarrow Fxyzw \)

where ∆_0 is consistency with P^0, ∆_1 is consistency with P^1, ∆_2 is consistency with P^2, and ∆_3 is consistency with P^3.

Thus (i) the modality of the quantifier is not univocal but depends upon its embedding, and (ii) the modality ∆_1 is determined holistically by P^1 on L^1. By "consistency with P^1" I mean merely that there are extensions L^1, formed by adding new names to L^1 whether or not those names have bearers, which extensions conform to (rl) to (r5).
(A): First we give the rules (r1) to (r5):-

(r1) I $P(A/B) \geq 0$
II $P(A/A) = 1$
III $(\exists A)(\exists B)(P(A/B) \neq 1)$
IV If $P(A/B) = 1 = P(B/A)$, then $(\forall C)P(C/A) = P(C/B)$
V $P(A \land B/C) = P(A/B \land C).P(B/C)$
VI $P(A \land B/C) \leq P(A/C)$
VII $P(\neg B/A) = 1 - P(B/A)$, unless $(\forall C)(P(C/A) = 1)$

(r2) For any variable $x$, for any formula $Fx$ of $L$ containing only $x$ free, and any singular terms $t_1, \ldots, t_n$ of $L$,

$$P(\exists x Fx/C) \geq P(F(t_1) \lor \ldots \lor F(t_n)/C)$$

(r3) For any variable $x$, for any formula $Fx$ of $L$ containing only $x$ free, for any sentence $C$ of $L$, and any number $q \leq P(\exists x Fx/C)$, there is an extension $L_+$ of $L$, and there are some terms $t_1, \ldots, t_n$ of $L$, and there is a reasonable extension $P_{+}$ of $P$ to $L_+$ such that

$$P_{+}(F(t_1) \lor \ldots \lor F(t_n)/C) \geq q$$

(r4) For any variable $x$, for any formula $Fx$ of $L$ containing only $x$ free, and any sentence $C$ of $L$, and any singular terms $t_1, \ldots, t_n$ of $L$,

$$P((\exists x Fx/C) \leq P(F(t_1) \land \ldots \land F(t_n)/C)$$

(r5) For any variable $x$, for any formula $Fx$ of $L$ containing only $x$ free, for any sentence $C$ of $L$, and any number $q \geq P((\exists x Fx/C)$, there is an extension $L_+$ of $L$, and there are $t_1, \ldots, t_n$ of $L$, and there is a reasonable extension $P_{+}$ of $P$ to $L_+$ such that

$$P_{+}(F(t_1) \land \ldots \land F(t_n)/C) \leq q$$

(r4) and (r5) are easily derivable from (r1) to (r3) using

$$P((\exists x Fx/C) = P(\neg (\exists x) Fx/C)$$

Df.

(B): We now suppose that all the sentences of $L$ are written in normal form with all the quantifiers gathered into a prefix, and we order the sentences of $L$ by the number of quantifiers in their prefixes: $L^0, L^1, L^2, \ldots, L^i, \ldots$. We then define $P$ on $L$ recursively as follows:

Base: Suppose $P^0$ on $L^0$ is reasonable (by r(1))

Induction step: Suppose $P^n$ is reasonable (by (r1) to (r5)) on $L^n$, and we define $P^{n+1}$ on $L^{n+1}$ as follows:-
There are eight cases to consider:

(i) \( P^{n+1}(\exists x)A/B \)
(ii) \( P^{n+1}((x)A/B) \)
(iii) \( P^{n+1}(A/(\exists x)B) \)
(iv) \( P^{n+1}(A/(x)B) \)
(v) \( P^{n+1}((\exists x)A/(x)B) \)
(vi) \( P^{n+1}((x)A/(\exists x)B) \)
(vii) \( P^{n+1}((x)A/(x)B) \)
(viii) \( P^{n+1}(\exists x)A/(\exists x)B \)

where \( A \) and \( B \) are closed sentences, or propositional functions, of order \( n \).

(i) \( P^{n+1}(\exists x)A/B \)

(a) For any extension \( L^n_+ \) of \( L^n \), for any terms \( t_1, \ldots, t_i \) of \( L^n_+ \), for any reasonable ((rl) to (r5)) extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \),

\[ P^{n+1}(\exists x)A/B \geq P^n_+(A(t_1) \ldots \land A(t_i))/B \]

(b) For any number \( q \), if \( q < P^{n+1}(\exists x)A/B \), then \( q < P^n_+(A(t_1) \ldots \land A(t_i))/B \) for some extension \( L^n_+ \) of \( L^n \) and some terms \( t_1, \ldots, t_i \) of \( L^n_+ \), and for some reasonable ((rl) to (r5)) extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \).

Here and in what follows, \( L^n_+ \) indicates an extension of \( L^n \) by adding new singular terms, whether or not they have bearers.

(ii) \( P^{n+1}((x)A/B) \)

(a) For any extension \( L^n_+ \) of \( L^n \), for any terms \( t_1, \ldots, t_i \) of \( L^n_+ \), for any reasonable ((rl) to (r5)) extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \),

\[ P^{n+1}((x)A/B) \leq P^n_+(A(t_1) \land \ldots \land A(t_i))/B \]

(b) For any number \( q \), if \( q > P^{n+1}((x)A/B) \), then \( q > P^n_+(A(t_1) \land \ldots \land A(t_i))/B \) for some extension \( L^n_+ \) of \( L^n \), and some terms \( t_1, \ldots, t_i \) of \( L^n_+ \), and for some reasonable ((rl) to (r5)) extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \).

(iii) \( P^{n+1}(A/(\exists x)B) \)

We want \( P^{n+1}(A/(\exists x)B) \) to take the minimum value of any \( P^n_+(A/B(t_1) \ldots \land B(t_n)) \) for any reasonable \( P^n_+ \) defined on any extension \( L^n_+ \). Hence:

(a) For any extension \( L^n_+ \) of \( L^n \), and any reasonable ((rl) to (r5)), extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \), for any terms \( t_1, \ldots, t_i \) of \( L^n_+ \),

\[ P^{n+1}(A/(\exists x)B) \leq P^n_+(A/B(t_1) \ldots \land B(t_i)) \]

(b) For any number \( q \), if \( q > P^{n+1}(A/(\exists x)B) \), then for some extension \( L^n_+ \) of \( L^n \), and some reasonable ((rl) to (r5)) extension \( P^n_+ \) of \( P^n \) to \( L^n_+ \), and some terms \( t_1, \ldots, t_i \) of \( L^n_+ \),

\[ q > P^n_+(A/B(t_1) \ldots \land B(t_i)) \]
200

(iv) \( p^{n+1}(A/(x)B) \)

The condition \((x)B\) is very strong: that every object in every possible extension satisfies \(B\). If this is not the case, then it is necessarily not the case and the condition is an impossible one, whence \( p^{n+1}(A/(x)B) = 1 \). If is is the case, then

\[
p^{n+1}(A/(x)B) = p^{n+1}(A)
\]

Hence:-

(a) IF for any extension \( L^+ \) of \( L^n \), and any reasonable \( ((r_1) \) to \( (r_5)) \) extension \( p^n \) of \( p^n \) to \( L^n \), and any terms \( t_1, \ldots, t_i \) of \( L^n \),

\[
p^*_n(B(t_1) \land \ldots \land B(t_i)) = 1,
\]

THEN

\[
p^{n+1}(A/(x)B) = p^n(A)
\]

(b) OTHERWISE, \( p^{n+1}(A/(x)B) = 1 \)

(v) \( p^{n+1}((3x)A/(x)B) \)

First we note that, by case (i), \( p^{n+1}((3x)A) \) is determined as follows:-

(a) For any extension \( L^+ \) of \( L^n \), for any terms \( t_1, \ldots, t_i \) of \( L^+ \),

for any reasonable \( ((r_1) \) to \( (r_5)) \) extension \( p^n \) of \( p^n \) to \( L^n \),

\[
p^{n+1}((3x)A) \geq p^n(A(t_1)v\ldots vA(t_i))
\]

(b) For any number \( q \), if \( q < p^{n+1}((3x)A) \), then \( q < p^n(A(t_1)v\ldots vA(t_i)) \) for some extension \( L^+ \) of \( L^n \) and some terms \( t_1, \ldots, t_i \) of \( L^+ \), and for some reasonable \( ((r_1) \) to \( (r_5)) \) extension \( p^n \) of \( p^n \) to \( L^n \).

Then we determine \( p^{n+1}((3x)A/(x)B) \) by case (iv) as follows:-

(c) If for any extension \( L^+ \) of \( L^n \), and any reasonable \( ((r_1) \) to \( (r_5)) \) extension \( p^n \) of \( p^n \) to \( L^n \), and for any terms \( t_1, \ldots, t_i \) of \( L^+ \),

\[
p^*_n(B(t_1) \land \ldots \land B(t_i)) = 1,
\]

THEN

\[
p^{n+1}((3x)A/(x)B) = p^{n+1}((3x)A)
\]

(d) OTHERWISE

\[
p^{n+1}((3x)A/(x)B) = 1
\]

(vi) \( p^{n+1}((x)A/(x)B) \)

This is determined in the same way as case (v), except that \( p^{n+1}((x)A) \) is first determined as in case (ii), and then \( p^{n+1}((x)A/(x)B) \) as in case (iv).
Obviously we want \( P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \) for any extension \( L^n_+ \) and any reasonable \( P^+_n \). Hence:

(a) For any extension \( L^n_+ \) of \( L^n \), and any reasonable ((r1) to (r5)) extension \( P^+_n \) of \( P^n \) to \( L^n_+ \), for any terms \( t_1, \ldots, t_i \) and any terms \( u_1, \ldots, u_j \) of \( L^n_+ \)

\[
P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \leq P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j))
\]

(b) For any number \( q \), if \( q \geq P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \), then for some extension \( L^n_+ \) of \( L^n \), and some reasonable ((r1) to (r5)) extension \( P^+_n \) of \( P^n \) to \( L^n_+ \), and some terms \( t_1, \ldots, t_i \) and \( u_1, \ldots, u_j \) of \( L^n_+ \)

\[
q > P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j))
\]

First we take some arbitrary extension \( L^n_+ \) and some arbitrary but reasonable \( P^+_n \) defined on it, and some terms \( u_1, \ldots, u_j \) from \( L^n_+ \), and define

\[
P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j))
\]

We want \( P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \) to take the maximum value of \( P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \) for any extension \( L^n_+ \) of \( L^n_+ \) and any reasonable ((r1) to (r5)) extension \( P^+_n \) of \( P^n \) to \( L^n_+ \). Hence:

(a) For any extension \( L^n_+ \) of \( L^n_+ \), for any terms \( t_1, \ldots, t_i \) of \( L^n_+ \), for any reasonable ((r1) to (r5)) extension \( P^+_n \) of \( P^n \) to \( L^n_+ \)

\[
P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \geq P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j))
\]

(b) For any number \( q \), if \( q < P_{n+1}(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \), then \( q < P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \) for some extension \( L^n_+ \) of \( L^n_+ \), and some terms \( t_1, \ldots, t_i \) of \( L^n_+ \), and some reasonable ((r1) to (r5)) extension \( P^+_n \) of \( P^n \) to \( L^n_+ \).

Secondly, we want \( P^n(A(t_1) \wedge \ldots \wedge A(t_i) / B(u_1) v \ldots v B(u_j)) \) for any \( L^n_+ \) and any reasonable \( P^+_n \) and any terms \( u_1, \ldots, u_j \) of \( L^n_+ \). Hence:
(c) For any extension $L_+^n$, any terms $u_1, \ldots, u_j$ of $L_+^n$, and any reasonable ((r1) to (r5)) extension $P_+^n$ of $P^n$ to $L_+^n$

$$P_{n+1}^+(\exists x)A/(\exists x)B \leq P_{n+1}^+(\exists x)A/B(u_1)\ldots vB(u_j))$$

(d) For any number $q$, if $q > P_{n+1}^+(\exists x)A/(\exists x)B$, then for some extension $L_+^n$ of $L^n$, and some terms $u_1, \ldots, u_j$ of $L_+^n$, and some reasonable ((r1) to (r5)) extension $P_{n+1}^+$ of $P^n$ to $L_+^n$

$$q > P_{n+1}^+(\exists x)A/B(u_1)\ldots vB(u_j))$$

In this way the original $P^0$ defined on $L^0$, the quantifier-free closed sentences of $L$, is extended recursively to all the closed sentences of $L$, and the resulting $P$ invites reading, for example

$$(\exists x)(y)(\exists z)F_{xyz}$$

as

$$\diamond_2(\exists x) \rightarrow \diamond_1(\exists y) \rightarrow \diamond_0(\exists z)F_{xyz}.$$
Appendix B to Chapter Four

Although it is not central to the question whether functionalism should be pitched wide or narrow, let me comment further on the problem syntactic inconsistency causes functionalists. Attitude reports exploit the inferential inter-relations between content clauses, and these are intended to map the inter-relations between the attitudes reported. Certain inferential relations are definitive of the senses of the logical constants (Hintikka: trivial inferences), and so if content clauses employing those logical constants map relations between z's states, it seems that we have the rule: If \( p \vdash q \), then \( z \) believes \( p \) \( \vdash p \). And hence the problem when we also want to say that Tom believes \( p \) but does not believe \( q \).

Perhaps the functionalist should seek a very weak set of inferential inter-relations to map a very minimal set of relations between psychological states? Loar takes this path, and we can follow his trail to expose the pitfalls.

Loar is a functionalist. He writes of beliefs:

- capturing the functional role of \( \ldots \) a belief does not require anything that seems to yield its truth conditions. The lateral connections with other beliefs and desires are what matter to its functional role, and not its vertical connections with the world.

In another work Loar gives rules which are intended to capture the functional roles of attitudes. Notably he gives rules connecting belief states, and these consist of rules connecting the ascription of belief with \( \text{not} \).
ascribing certain other beliefs. Here are some examples, reading "zBp" as z believes that p:-28

1(a) Not zB-(pv-p)
1(b) If zBp, then not zB-p
2(a) If zBp&q, then not zB-p and not zB-q
4(a) If zBp→q and zBp, then not zB-q

Loar remarks29:-

if any of them is not by and large true of z, beliefs of the relevant logical form do not seem ascribable to z; the required rationality is so minimal that their approximate satisfaction is partially definitive of having the relevant beliefs.

Thus Loar's inferential rules make very little demand on z's logical acumen. They forbid certain beliefs, and go only from having certain beliefs to not having certain other beliefs. And the requirements are further weakened because z is only required to adhere to them "by and large".

It seems to me that this approach encounters two insuperable obstacles. Firstly, Loar claims that his rules are only obeyed "by and large", but also claims that they partially define the functional roles which are those belief states. If the rules are only obeyed by and large then this implies that a believer could have a certain specific belief even though the relevant rules are broken on this particular occasion. But this is impossible on Loar's theory: a believer cannot be in a functional state unless that state has its defining functional relations.

Loar has given us no independent classification of beliefs to enable us even to make sense of the idea of someone having a given belief which fails to have its usual functional relations.
Secondly, in Loar's system, since all the rules go from belief to lack of belief, no chains of consequential belief need be built up. For example, the following triad is consistent with those rules:

\[ zBp \land q, \ zBp \rightarrow r, \ zB \lnot r \]

For from the rule 2(a) and \( zBp \land q \) we will be able to conclude only

\[ \lnot zB \lnot p \]

and from \( zBp \rightarrow r \) and \( zB \lnot r \) and rule 4(a) we can conclude only

\[ \lnot zBp \]

which are mutually consistent. Hence, even if we take strict, not partial, adherence to Loar's rules as definitive of the relevant functional states, then the functional theory could be satisfied, in principle, by a bizarrely irrational creature: its behaviour need show no attunement to its environment, no goal-directed success, since no chains of true beliefs causally consequent upon some initial true beliefs need ever be set up. Loar's rules are too weak to define functional states which mediate adaptive behaviour.

Weakening the inferential rules does not help the functionalist. In my view, functionalists should employ the strong rule: If \( p \vdash q \), then \( z \) believes \( p \vdash z \) believes \( q \), where \( \vdash \), we should note, is syntactic implication not semantic implication - Pierre's beliefs, as we saw in Appendix B to Chapter Two, are semantically inconsistent but not syntactically inconsistent. Seeking a minimal standard of rationality below which we cannot fall is
pursuing a Will o' the Wisp. There need be no inference
p \rightarrow q so trivial that no-one anywhere has failed to draw it,
or any precise boundary between those we generally get
right and those we often get wrong. Rationality is a
matter of being right often enough. But there need be no
lowest common denominator of cases we all get right.

But, now, how do we reconcile such a strong functional
theory with the datum that we sometimes believe that p,
but not that q, when p \rightarrow q? I suggest we do not try: when
this happens we have a malfunction, we no longer, or do
not in that respect or for that episode, realise the
requisite physical and psychological functional theory.
The explanation of a malfunction lies outwith the functional
theory. Given a machine which sometimes works and sometimes
doesn't, the functional theory need not predict or explain
the malfunctions of the machine.

When folk-psychology declares that a subject fails to
believe some syntactic (not semantic - remember Pierre)
consequence of a given belief, then our functionalist should
abandon any claim to give a global functional explanation
spanning both beliefs. Rather, the subject's behaviour
will divide into two subclasses of behaviour, one of which
will be (partially) explained by the given belief taken to
be a causally active state mediating between perceptual
input and this subset of behavioural output, and by belief
in all the consequences of that belief, and the other
subdivision of the subject's behaviour will be (partially)
explained by the lack of the other belief and the lack of
any belief which would imply it. So long as we take concrete attributions of belief to a subject to be relative to some restricted explanatory context, some particular behavioural output, the global attribution of syntactically inconsistent beliefs does not threaten functionalism.

Belief explains behaviour under its intentional description - raising the arm rather than the arm rising. Also, one and the same bodily motion may be the performance of two distinctly motivated acts - as by whispering loudly I may both intentionally greet a friend and intentionally irritate the visiting speaker. So it may even be that one and the same bodily movement is explained under the intentional description d by the belief that p, and under the intentional description d' by the belief that q, where p and q are syntactically inconsistent. I.e., one and the same bodily motion can be two different actions each belonging to a different functional sub-division of the subject's total behaviour. No wonder folk psychology declares us to have syntactically inconsistent beliefs.

What about the subject who passes the general test for understanding both p and q, who endorses p and then carries out Hintikka's canonical test to determine whether p trivially implies q, makes a slip and denies q? If we explain his endorsing p by a belief that p, then that act is also explained by his belief that q, and in all other consequences of p. If we explain his later act of denying that q by his not believing that q, then that act is also explained by his not believing that p, nor any other
proposition that implies q. There is no functional explanation of the two acts together, because his doing both was a malfunction.

The above does not mean that there was not one thing it was like to be that person, that he was in two minds. It means, rather, that the one thing it was like to be that person is disjointed to the extent that his beliefs were syntactically inconsistent. To fail the criterion of psychological realism the/would have to be a second psychological history, whether disjointed or not, of the same behavioural sequence.
Footnotes to Chapter Four

1 Fodor (2) p108
2 Fodor (2) p101, see also the exchange with Haugeland.
3 Fodor (2) p105
4 Fodor (3) p208
5 Fodor (3) p209
6 Fodor (3) p238 fnnte 5
7 Searle(1) provides an example which, to my mind, makes the same point as this section - viz. that those narrow functional relations which realise the machine-code computations, do not thereby realise the concepts and cogitations which the agent grasps and exercises. Searle imagines a man in a room armed with a program, written in English, who receives input data in the form of sentences written in Chinese and who follows the program to produce output also in the form of sentences written in Chinese: the man understands only English, and his program instructs him to match Chinese characters to Chinese characters, store them in boxes, move them from box to box, etc. The input and output sentences together form intelligent questions and answers in Chinese about some story. One of Searle's points is that the man in the room does not understand the Chinese sentences he receives and produces, although he does fully understand the 'machine code' English of the program he uses. And nothing he does understand is equivalent to the senses of the Chinese sentences, so we cannot say that he really understands the Chinese sentences, but without realising the fact.

Searle's view of what more is required to understand Chinese is somewhat obscure to me. For he also holds, in Searle (2) that all which determines a Chinaperson's conceptual grasp is contained within the neural envelope.

8 Field (3)
9 Hintikka "Surface Information and Depth Information" in Hintikka and Suppes (1), and Hintikka "Impossible
Worlds Vindicated" in Saarinen (1)

10 Rantala "Urn Models: A new kind of non-standard model for first-order logic" in Saarinen (1)

11 McGinn "The Structure of Content" in Woodfield (1)

p234

12 Field (3) p379

13 Field (3) p383

14 Field (3) p387

15 Field (3) p386

16 Field (3) p386 ft 10

17 It is important to read ...(x)...Fx as \( \neg \Diamond_1(\exists x)\neg \ldots \neg Fx \) and not as \( \ldots \Diamond_1(\exists x)\neg \ldots \neg Fx \). For on the latter reading, \((x)Fx\) and \((\exists x)\neg Fx\), for example, would become consistent, being \( \Diamond_1(\exists x)\neg Fx \) and \( \Diamond_1(\exists x)\neg Fx \), and so the belief-system \( P \) would become classically inconsistent, since \((x)Fx\), \((\exists x)\neg Fx\) \( \vdash \mathrm{A} \land \neg \mathrm{A} \). But on the first pair of readings \((x)Fx\) and \((\exists x)\neg Fx\) are inconsistent, as required, being \( \neg \Diamond_1(\exists x)\neg Fx \) and \( \Diamond_1(\exists x)\neg Fx \).

18 This is an allusion to Hintikka's explanation of the meanings of the quantifiers using seeking and finding games, in Hintikka (2)

19 Hintikka "Surface Information and Depth Information" in Hintikka and Suppes (1) p294

20 Hintikka "Impossible Worlds Vindicated" and Rantala "Urn Models: A new kind of semantics for first-order logic", both in Saarinen (1)

21 Details in Hintikka (2)

22 Saarinen (1) p371

23 Saarinen (1) p371/2

24 Saarinan (1) p372

25 Saarinen (1) p356

26 Fodor (2) p108

27 In Mellor (1) p67

28 Loar (1) p72

29 Loar (1) p71
Chapter Five: De re attitudes and psychological explanations

Section (i): Introduction

I shall argue that de re attitude reports play an essential role in psychological explanations, and that they report de re attitudes whose res are the entities which the psychological subject perceives and manipulates in his wide functional system.

Dennett\(^1\), Fodor\(^2\), and McGinn\(^3\) argue that de re reports of attitudes have no part to play in psychological explanations, and that the de re attitudes reported are not essential to the psychological processes. Dennett makes a brief and forceful case for the explanatory impotence of de re attitude reports when he writes\(^4\):-

we would not suppose that anything at all about Tom's likely behaviour follows from the truth of Tom believes of the man shaking hands with him that he is a heavily armed fugitive mass murderer.

This view acquiesces in what we might call the psychological opacity of semantic transparency ..., and while I can see no obstacle to defining such a variety of propositional attitude, I can see no use for such a concept, since nothing of interest would seem to follow from the true attribution of such a belief. Suppose, on this view, that someone, a, is believed by me to be F. It does not follow that a is not also believed by me to be not-F; and if a is also believed by me to be G, it does not follow that a is believed by me to be F and G, it does not even follow from the fact that a is believed by me to be the only F, that no other object b is also believed by me to be the only F.

And Dennett immediately continues by conferring psychological opacity on the attitudes so reported:-

The premise for the quest for de re belief was that there were interesting and important relations between believers and the objects of their beliefs - relations we had reasons to capture in our theories - but this termination of the quest lands us with relations that are of only intermittent and unprojectible interest.
McGinn agrees with Dennett that:

the explanatory force of the content ascription [of a belief report] attaches only to the contribution the words in the content clause make in their capacity as specifiers of internal representations; their referential properties play no explanatory role.

McGinn proposes:

that we locate reference in the point of communication - in the intentions with which assertions (and other kinds of speech act) are made. A hearer understands a speech act as an assertion just if he interprets it as performed with a certain point or intention - viz. to convey information about the world ... we will ... need semantic concepts to explain why it is that the speaker chose those words to convey knowledge (or belief) about a certain state of affairs to an audience.

So it seems that we do need semantic concepts to explain the speaker's intentions in choosing the words he did to convey knowledge or belief about the world to an audience.

However, McGinn's view seems to be that in specifying and explaining that intention we do not need to assign specific referents to the speaker's internal representations. McGinn writes:

It should be noted that my suggestion is not that the reference relation itself explains the speaker's act of communication: I am saying rather that the thought of it is presupposed in reflective mastery of the practice, or in characterising its function where the practice seems quite unreflective.

Fortunately, McGinn gives an example elsewhere which illuminates this last oracular pronouncement. McGinn considers, as a test case, the reader trying to raise his arm, and concludes:

you do try to raise your arm, but its being that arm - a particular physical object outside the mind - is immaterial to what you try; you try, we might say, to raise whatever physical arm is appended to your will in the appropriate way. This is not to deny that your trying is directed at a particular limb; it is so
directed, but only in the way that perceptual experiences are directed at particular external objects: there is a notion of content for both which is independent of the identity of these objects and even of their existence.

As we shall see in the next chapter, I also think that to explain the agent's behaviour, to explain why he chose those words to fulfil his intention to communicate, say, we do not need to attribute specific res to his attitudes. The explanation will quantify over the res: the explanation is 'blind to' the particular values of the variables which make the quantified sentences true. The functional explanation abstracts from the particular identities of the res involved. But McGinn goes further: his quantifier lies within the scope of the psychological verb - you try to raise whatever arm is appropriately appended to your will - hence the explanation is independent of even the existence of an arm.

Fodor too argues that de re attitude reports have no part to play in psychological explanations. He recommends what he calls rational psychology and writes:

so far as rational psychology per se is concerned, S [the subject] could be an ideal entity "with no links to the physical world". That is: whether the properties that the rational psychologist attributes to a mental state are in fact properties of that state isn't allowed to depend essentially on the existence of such links.

Again, note, that Fodor is not claiming merely, as I shall in the next chapter, that psychological identities do not depend essentially upon the particular identities of any res. He is claiming that they do not depend essentially upon the existence of any res. Whence psychological
explanations do not employ de re reports.

To oppose this received wisdom I shall set out my alternative account of the psychological determinants of behaviour, and of psychological explanation, in section (ii). In section (iii) I will support my account by linking it to the theory of meaning, and in section (iv) give an account of the satisfaction of quantified de re reports by sequences. In section (v) I consider what must seem the most damaging objection to a theory which gives an essential role to de re attitudes, viz. hallucinations. And finally, in section (vi) I consider whether 'modes of presentation' usurp the role I find for de re attitudes in determining action.

Section (ii): The role of de re attitude reports in psychological explanations. And the role of de re attitudes in the causation of action

As I have already partially explained in Appendix B to Chapter Two, de re attitude reports are required to link up and mutually orient reports of perceptions, beliefs, desires and intentions into an explanatory complex, explaining some intentional action or some change of belief. Further, if we are to have a psychological theory, as well as individual psychological histories, we will need general functional laws relating beliefs, desires, etc. in rational patterns. Such laws, I shall argue, employ de re attitude reports. So my immediate answer to the Dennett passage quoted above is that from only one de re attitude report of the form
we learn nothing of Tom's likely behaviour. But a different picture emerges when we consider \textit{de re} reports structured as follows:--

\begin{equation}
(\exists x_1 x_2 x_3 x_4)(\text{Tom perceives } [Fx_1] \land \text{Tom believes } \left[ Gx_2 \land x_1 = x_2 = x_3 = x_4 \right] \land \text{Tom desires } [Hx_3] \land \text{Tom intends } [Ix_4])
\end{equation}

(Note that only in the identity conjuncts of the belief clause do the variables used in the other attitude reports recur. This is intended to allow the maximum scope for recording Tom's beliefs as to the identities and differences among the various \textit{res}.) Further, the bound variables mutually orient the attitudes within the scope of the quantifiers. For example

\begin{equation}
(\exists x_1 x_2 x_3 x_4)(\text{Tom bel}[Fx_1 x_2 \land x_1 = x_3 \land x_2 = x_4] \land \\
\text{Tom des } [Fx_3 x_4])
\end{equation}

suggests Tom will be quiescent, whereas

\begin{equation}
(\exists x_1 x_2 x_3 x_4)(\text{Tom bel}[Fx_1 x_2 \land x_1 = x_4 \land x_2 = x_3] \land \\
\text{Tom des}[Fx_3 x_4])
\end{equation}

suggests Tom will be rearranging the furniture of the universe. (Let me use "bel" for "believes", "per" for "perceives", etc. And "[ ]" indicates the scope of a psychological verb.)

However, it is one thing to claim that our psychological explanations need to employ \textit{de re} reports, and another to claim that the attitudes reported are themselves \textit{de re}. My view is that, in general, they are: that feature of our reports reflects a feature of the psychological facts reported.

Consider, for example, a detective working on a murder.
He sees a footprint in the flower bed below the library window, then another, and another leading towards the scullery door. He comes to the conclusion that the murderer left the library by the window and rejoined the other guests by re-entering the house through the scullery. Suppose our man in the trenchcoat successively discovers some two hundred or so prints, successively forming beliefs evinced by such utterances as "That's a footprint", "That's another footprint", "And another", etc. He comes to believe that there is a continuous trail of footprints leading from library to scullery, and the folk-explanation of how he comes to this belief seems to require

$$(\exists x_1 \ldots x_n) \left( \text{D sees} [Fx_1] \land \text{D sees} [Fx_2 \land x_2 \neq x_1] \land \ldots \land \text{D sees} [Fx_n \land x_n \neq x_1 \land x_2 \land \ldots \land x_{n-1}] \right)$$

where D is the detective and Fx is "x is a footprint".

Sometimes, when the detective observes a footprint at $t_1$, turns to speak to his sergeant, and returns at $t_2$ to contemplate the same footprint again, we have:

$$(\exists x_1 x_2) \left( \text{D sees at } t_1 \text{ that } [Fx_1] \land \text{D sees at } t_2 \text{ that } [Fx_2] \land \text{D believes at } t_2 \text{ that } [x_1 = x_2] \right)$$

Psychological explanations seem thus to employ de re reports of de re attitudes.

My view is that these de re reports do correctly report de re attitudes. The detective's judgements of identity and difference are primitive, so far as his conscious judgements are concerned. Admittedly, the detective cannot be aware of a footprint without conceptualising it in some way or other. But, I contend, his conscious conceptualisation of the $i$th footprint may be
exactly the same as his later conscious conceptualisation of the $j$th footprint except that he conceives of them as being different footprints.

Of course, we expect a computational psychology to explain such (at the conscious level) primitive judgements of identity and difference, but that explanation need not import further content into his conscious judgements. That explanation may be a 'machine code' computation: operations defined on the 'syntax' of the detective's internal representational states, but operations which do not add conceptual content to those representational states.

Taking the representational content of an internal state to be determined by its functional relations in a (wide) functional system, it seems those functional relations are supposedly able to bestow upon a given state the representational content of something being a footprint. But if functional relations can bestow the content of being-a-footprint on an internal state, they can surely bestow being-the-same-footprint-as or being-a-different-footprint-from that represented by one (or more) of the other related states.

If it is regarded as unproblematic that functional relations are able to bestow that commonality of representational content which is reported by the same predicate expression in the content clause, then it is equally unproblematic, surely, that those relations can bestow that commonality of content which is reported by a commonly bound variable. Nothing easier, it would seem, than to make an internal state with the representational
content of being an F causally beget another also having
the representational content of being an F. Why not, then,
one with the content of being the same (or a different) F? 10

Suppose, where \( t_1 \) and \( t_2 \) are successive times of
footprint sightings, we have:-
\[
(\exists x_1 x_2)(D \text{ per-at-} t_1[Fx_1] \& D \text{ per-at-} t_2[Fx_2] \&
D \text{ bel-at-} t_2[Fx_1 \& Fx_2 \& x_1 \neq x_2])
\]
Or suppose we have:-
\[
(\exists x_1 x_2)(D \text{ per-at-} t_1[Fx_1] \& D \text{ per-at-} t_2[Fx_2] \&
D \text{ bel-at-} t_2[Fx_1 \& Fx_2 \& x_1 = x_2])
\]
Suppose there is no further conscious conceptualisation of
the footprints, no common or different modes of presentation
or 'vivid names', which warrant the detective's judgement
that this is a different, or that this is the same footprint.
Does this mean that his judgement is unwarranted, irrational?

Internal 'machine code' processes give the detective's
successive representational states the contents reported,
'bare' judgements of identity or difference, we suppose.
And the machine code computation may warrant the judgement
of identity or difference because it preserves the appropriate
semantic value. However, this is not a warrant which is
consciously available to the believer. From the believer's
conscious point of view, the warrant for such 'bare'
judgements of identity and difference is holistic: it lies
in the rationality of the total belief system they in part
make up. What consciously warrants the detective's
judgement that those sightings were sightings of the same
footprint, whilst these sightings were sightings of a
different footprint, is the coherence of the resulting
belief system, it being a better explanation of the world
around him than one incorporating the reverse judgements. The conscious warrant for the judgement of identity or difference is the success of the resulting theory.

My reader may feel that whilst quantification-in may well bind and mutually orient attitude reports, the resulting bonds are far too powerful, generating referentially transparent contexts, to be of use in psychological explanations. Suppose for example, we have:

\[(\exists x_1 x_2)(x_1 = x_2 \& \text{Tom bel}[F_{x_1} \& x_1 = x_2 \& F_{x_2}]) \quad \ldots (1)\]

Does this not imply

\[(\exists x)(\text{Tom bel}[F_x \& x \neq x \& F_{\neg x}]) \quad \ldots \ldots (2)\]

which, being syntactically inconsistent, does not index any belief state? We avoid this implication by employing a paratactic analysis and the quasi-binding of variables within the demonstrated clause. Thus (1) becomes:

\[(\exists x_1 x_2)(x_1 = x_2 \& \text{Tom bel of } \langle x_1 x_2 \rangle \text{that } [F_{x_1} \& x_1 = x_2 \& \neg F_{x_2}]) \quad (1')\]

where the variables in \(\langle \rangle\) are transparent, but those in \([\phantom{\text{dummy}}]\) are only quasi-bound. This means that (1') implies only

\[(\exists x)(\text{Tom bel of } \langle x_1 x_2 \rangle \text{that } [F_{x_1} \& x_1 = x_2 \& \neg F_{x_2}]) \quad \ldots (3)\]

where the initial quantifier binds both the variables in \(\langle \rangle\), and thereby quasi-binds the variables in \([\phantom{\text{dummy}}]\). But it only quasi-binds those variables: the "x_1" and "x_2" in \([\phantom{\text{dummy}}]\) are different variables, and the subscripts indicate the different places in \(\langle \rangle\) they are bound to, whereas "x_1", "x_2" in \(\langle \rangle\) are two occurrences of the same variable.

On the other hand, (2) is

\[(\exists x)(\text{Tom bel of } \langle x_1 \rangle \text{ that } [F_{x_1} \& x_1 \neq x_1 \& \neg F_{x_1}]) \quad (2')\]

a psychological state no-one is in.
However, it may still seem that quantification-in is too transparent for the purposes of psychological explanation. Suppose we have a psychologically explanatory complex of the form:

\[(\exists x_1 x_2 x_3 x_4) (\text{Tom per of } x_1 \text{ that}[F_{x_1}] \land \text{Tom bel of } x_1 x_2 x_3 x_4 \text{ that}[G_{x_2} \& x_1 \neq x_2 \neq x_3 \neq x_4] \land \text{Tom des of } x_2 \text{ that}[H_{x_3}] \land \text{Tom int of } x_4 \text{ that}[J_{x_4}]) \]  

\[\] 

(4)

Now suppose in fact, the relevant $x_1 = x_2 = x_3 = x_4$. Then we have:

\[(\exists x) (\text{Tom per of } x_1 \text{ that}[F_{x_1}] \land \text{Tom bel of } x_1 x_2 x_3 x_4 \text{ that}[G_{x_2} \& x_1 \neq x_2 \neq x_3 \neq x_4] \land \text{Tom des of } x_2 \text{ that}[H_{x_3}] \land \text{Tom int of } x_4 \text{ that}[J_{x_4}]) \]  

\[\] 

(5)

But (5) is only notionally different from

\[(\exists x) (\text{Tom per of } x_1 \text{ that}[F_{x_1}] \land \text{Tom bel of } x_1 x_2 x_3 x_4 \text{ that}[G_{x_1} \& x_1 \neq x_2 \neq x_3 \neq x_4] \land \text{Tom des of } x_1 \text{ that}[H_{x_1}] \land \text{Tom int of } x_1 \text{ that}[J_{x_1}]) \]  

\[\] 

(6)

for the subscripts only relate the quasi-bound variables in the \[\] clauses to the members of their respective sequences \(<\>\). If two occurrences of \(<\> contain the same variables, the use of different numerical subscripts to link them to the variables they quasi-bind in their respective \[\] clauses is of no significance. But (6), which seems equivalent to (5), given $x_1 = x_2 = x_3 = x_4$, contains a different linking up and mutual orienting of the paratactic clauses. If (5) is an explanatory complex, we have no reason to expect (6) to be one also. So, it seems, the binding and quasi-binding of variables across content clauses is too strong for our purposes of forming explanatory complexes.

However, I contend that, assuming $x_1 = x_2 = x_3 = x_4$, (5) and
are not equivalent; that whilst (5) implies (6), (6) does not imply (5). (6) does not imply (5) because the content clauses in (5) are linked up and mutually oriented, but not so those in (6). (6) tells us nothing about how Tom thinks of the res of his various attitudes in relation one to another. (6) does not even require that Tom thinks of the attitudes as having any res in common at all. To link up and mutually orient individual attitudes into an explanatory complex, we need a quantifier prefix "(∃x₁...xₙ)" such that:

(i) Wherever in a content clause [ ] we have different variables xᵢ,xⱼ, then they are bound by different variables xᵢ,xⱼ in the sequence < >, which are themselves bound by different quantifiers ∃xᵢ,∃xⱼ.

(ii) No variables are common to more than one attitude clause, except that all the variables recur in the belief clause in sub-clauses of the form xᵢ=xₙ or xₙ≠xₙ.

We can then take the variables in the various clauses to show how those thoughts are linked up and mutually oriented by the agent.

There is no a priori rule for going from individual attitude reports, say

(∃x₁...xₙ)(Tom bel of <x₁...xₙ> that[Fx₁...xₙ])
(∃x₁...xₙ)(Tom des of <x₁...xₙ> that[Gx₁...xₙ])
(∃x₁...xₙ)(Tom int of <x₁...xₙ> that[Hx₁...xₙ])

to an explanatory complex. Given the individual attitude reports, we do not know from the variables employed in them, whether or not Tom links the thoughts up, nor, if he does, how. We infer explanatory complexes from
individual attitude reports on the holistic grounds that the resulting explanation fits Tom's behaviour, including verbal behaviour of course. I.e., the resulting explanatory complex fits the psychological functional theory of the behaviour of Tom and his fellow language-speakers.

Explanation is relative to theory, and explanatory complexes contain reports of specific psychological states which exemplify general psychological functional laws. I take the psychological theory of Tom and his peers to include the following laws (A) to (E).

(A) A general rule linking assertion and belief - as discussed in Chapter Two section (vi):-

\[
(x)(p)\{Nx \rightarrow (x \text{ believes } p) \iff (\exists t)(\text{Axt} \& s(p) = s(t))
\]

& (Js)(P_{ps} \vdash P_{ts})

where "Nx" says that x is a normal English speaker, "Axt" says that x is disposed to give sincere reflective assent to a token t, "s(p) = s(t)" requires that the tokens p and t have the same semantic content, and "(Js)(P_{ps} \vdash P_{ts})" requires that there is some uniform subscripting of the singular terms in p and in t which gives the same subscript to a term in p and in t only if they have the same referent, and does not give the same subscript to different singular terms in p or in t (reflexive pronouns, a second occurrence of a proper name are recurrences of the same singular term), and such that, so subscripted $P_{ps}$ and $P_{ts}$ are syntactically interderivable.

(B) We need a rule for moving from "z believes p", where p contains $a_1, \ldots, a_n$ distinct singular terms, to "(\exists x_1 \ldots x_n)z \text{ believes } p_{x_1/a_1} \ldots p_{x_n/a_n}". (I am ignoring the
complication introduced by McGinn's point, discussed in Appendix C to Chapter Two, that acts of reference need not manifest a belief de the referent. Either we can hope to build that complication into (A) above and have an uncomplicated rule (B), or we can insert whatever conditions into (B) we think necessary for de re belief.

$$(z)(\forall \text{bel}[P_{a_1 \ldots a_n}] \Rightarrow (3x_1 \ldots x_n) (x_1 = a_1 \& \ldots \& x_n = a_n \\
\& \ z \ \text{bel} \ of \ <x_1 \ldots x_n> \ \text{that}[P_{x_1/a_1 \ldots x_n/a_n}])$$

And clearly, if we can quantify over all the singular terms in "Pa_1 \ldots a_n", we also want to be able to quantify over just some of them.

When we seek to explain, holistically, how some particular beliefs lead to other beliefs, or how beliefs and desires lead to actions, we need the quantifiers and the variables they quasi-bind to link up and orient the various attitude reports. Here the subscripting of variables in $<$ and in $]$ in one component must not only show which members of $<$ are the values of which variables in $]$, but the subscripting must be common to all the components to provide the required links and orientation.

(C) Suppose $P_{x_1 \ldots x_j} \vdash Q_{x_k \ldots x_L}$, where $x_1, \ldots, x_j$ are all the free variables in $P$ and $x_k, \ldots, x_L$ are those in $Q$, and $\vdash$ is obviously a logic which can handle open sentences. Let $\{x_1, \ldots, x_n\}$ be the total set of variables

$$\{x_1, \ldots, x_j\} \cup \{x_k, \ldots, x_L\}$$

Then:

$$(z)(x_1 \ldots x_n)(z \text{bel} \ of \ <x_1 \ldots x_j> \ \text{that}[P_{x_1 \ldots x_j}]) \Rightarrow \ z \ \text{bel} \ of \ <x_k \ldots x_L> \ \text{that}[Q_{x_k \ldots x_L}])$$

(D) Something along the lines of a practical syllogism.
Suppose $F_{x_1,...,x_n}, G_{x_1,...,x_j} \vdash H_{x_k,...,x_L}$ where $G_{x_1,...,x_j}$ and $H_{x_k,...,x_L}$ have no variables in common, and where $\{x_1,...,x_n\}$ includes both $\{x_1,...,x_j\}$ and $\{x_k,...,x_L\}$, and may have extra variables of its own. (Since $\{x_1,...,x_j\}$ and $\{x_k,...,x_L\}$ are disjoint, we expect $F_{x_1,...,x_n}$ to include clauses of the forms $x_i=x_k$ and $x_j=x_L$ and the like. Then:

$$(z)(x_1,...,x_n)((z \text{ bel of } \langle x_1,...,x_n \rangle \text{ that } [F_{x_1,...,x_n}]
\& \ z \text{ des of } \langle x_k,...,x_L \rangle \text{ that } [H_{x_k,...,x_L}]) \rightarrow
z \text{ int of } \langle x_1,...,x_j \rangle \text{ that } [G_{x_1,...,x_j}])$$

In addition to rules (A) to (D), true of all subjects, we will expect rules which give the different castes of mind of different subjects - who therefore realise the same generic psychological functional theory, but different specific psychological functional theories. For example, jealous Jack, if he believes that some man admires Jill, then he fears that Jill admires that man.

$$(E) (x_1 \times x_2)(\text{Jack believes of } \langle x_1 x_2 \rangle \text{ that } [\text{Male } x_1 \& \text{ Admires } x_1, \text{Jill } & x_1=x_2] \rightarrow \text{Jack fears of } \langle x_2 \rangle \text{ that } [\text{Admires Jill, } x_2])$$

The point to note about these rules emerges in (C) and (D). They exploit syntactic relations $\vdash$ between the paratactically denoted clauses. Quantification-in, and the quasi-binding of variables in the paratactically denoted clauses, are needed to set up those syntactic deducibility relations. Quantification-in does reach the variables in the paratactical clauses, otherwise we would not have common variables in the clauses $P, Q, R,...,S$ so as to permit $P, Q, R,... \vdash S$. But quantification only quasi-binds
those variables, in the sense that where we have, e.g.,
\((\exists x_1 x_2)(x_1=x_2 \land \ldots)\) we may not reduce the number of
variables quasi-bound in any \([\ ]\), but only in their
attendant \(<\ >\text{terms}.

These features of the psychological functional theory
motivate the prohibitions of which Dennett complains\(^4\). We
should not expect, for example:–

\[(\exists x_1 x_2)(x_1=x_2 \land \text{Tom bel}[Fx_1 x_2]) \vdash (\exists x)(\text{Tom bel}[Fxx])\]

nor

\[(\exists x)(x=a \land \text{Tom bel}[Fx]) \vdash \text{Tom bel}[Fa]\]

since these are, respectively, the invalid:–

\[(\exists x_1 x_2)(x_1=x_2 \land \text{Tom bel of} <x_1 x_2> \text{that} [Fx_1 x_2])\]

\[\# (\exists x)(\text{Tom bel of} <x_1> \text{that} [Fx_1 x_1])\]

\[(\exists x)(x=a \land \text{Tom bel of} <x_1> \text{that} [Fx_1])\]

\[\# \text{Tom bel of} <a_1> \text{that} [Fa_1]\]

The last are invalid because we do not have \(F_{x_1 x_2} \vdash F_{x_1 x_1}\)

nor \(F_{x_1} \vdash Fa_1\) as rule (C) above would require.

Dennett specifically complains that, given,

\[(\exists x)(\text{Tom bel} [Fx] \land \text{Tom bel} [Gx]) \quad \ldots \ldots \ldots \ldots (7)\]

we may not infer

\[(\exists x)(\text{Tom bel} [Fx \land Gx]) \quad \ldots \ldots \ldots \ldots (8)\]

and that we can have

\[(\exists x)(\text{Tom bel} [Fx \land \neg Fx]) \quad \ldots \ldots \ldots \ldots (9)\]

I reply that, reading (8) as

\[(\exists x)(\text{Tom bel of} <x_1> \text{that} [Fx_1 \land Gx_1]) \quad \ldots \ldots (8')\]

it is implied by (7), if (7) is read as

\[(\exists x_1 x_2)(\text{Tom bel of} <x_1 x_2> \text{that} [Fx_1 \land x_1=x_2]
\land \text{Tom bel of} <x_2> \text{that} [Gx_2]) \quad \ldots \ldots (7')\]

since \(Fx_1 \land x_1=x_2, Gx_2 \vdash Fx_1 \land Gx_1\)
but not if (7) is read as

\[(\exists x_1 x_2) (\text{Tom}\ \text{bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 \& x_1 \neq x_2] \& \text{Tom}\ \text{bel of } \langle x_2 \rangle \text{ that } [Gx_2]) \] ........................(7'')

nor if (7) is read as

\[(\exists x_1 x_2) (x_1 = x_2 \& \text{Tom}\ \text{bel of } \langle x_1 \rangle \text{ that } [Fx_1] \& \text{Tom}\ \text{bel of } \langle x_2 \rangle \text{ that } [Fx_2]) \] ........................(7'''')

And (9) is not a possible belief if we read it as

\[(\exists x) (\text{Tom}\ \text{bel of } \langle x_1 \rangle \text{ that } [Fx_1 \& -Fx_1]) \] ........................(9')

but it is a possible belief read as

\[(\exists x) (\text{Tom}\ \text{bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 \& -Fx_2]) \] ........................(9'')

or as the equivalent

\[(\exists x_1 x_2) (x_1 = x_2 \& \text{Tom}\ \text{bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 \& -Fx_2]) \] ........................(9'''')

Thus, I claim, the way in which de re attitudes mesh into a functional theory explains the prohibitions which Dennett wrongly thinks renders them useless in psychological explanations ("psychologically opaque"): the key is quantification-in plus a paratactic reading to yield quasi-bound variables.11

Section (iii): Utterance meaning and the diversity of individual functional psychological theories

In this section I wish to show how a psychological functional theory, as sketched in the last section, connects with a theory of the meanings of utterances. And connects in such a way that the common psychological theory of the utterers incorporates plausible individual differences between the functional inter-reations of one utterer's psychological states and the functional inter-relations of another utterer's psychological states, though both
grasp the same utterance meanings.

Consider first a language containing no indexical elements, whose only singular terms are context free proper names. (In such a language, of course, there is no distinction between sentence meaning and utterer's meaning.) This restriction will be lifted later. A theory of utterance meaning for such a language should recursively assign to each token utterance \(P\) an ordered pair \(\langle TC_P, S_P \rangle\), where \(TC_P\) are the truth conditions of \(P\) and \(S_P\) is the syntactic structure of \(P\).

I.e., \(M(P) = \langle TC_P, S_P \rangle\)

And this assignment is to be recursive in that if \(F\) is a syntactic one-place sentence forming operator of the language, the theory associates with \(F\) a pair of functions \(F_1\) and \(F_2\) such that

\[
M(F(P)) = \langle F_1(M(P)), F_2(M(P)) \rangle
\]

where \(F_1\) is a uniform function from \(M(P)\) to the truth conditions of \(F(P)\), and \(F_2\) is a uniform function from \(M(P)\) to the syntactic structure of \(F(P)\).

We also require that the theory of meaning retain the following link with a Fregean conception of sense: If it is possible for an utterer to believe that \(P_1\) but not \(P_2\), or vice versa, and if \(TC_{P1} = TC_{P2}\), then \(S_{P1} \neq S_{P2}\). I.e., an utterer can believe \(P_1\) but not \(P_2\), so they differ in meaning, but they do not differ in truth conditions, and so must differ in syntactic structure. And, vice versa, if \(TC_{P1} = TC_{P2}\) and \(S_{P1} \neq S_{P2}\), then it is possible for a believer to believe \(P_1\) and not \(P_2\), or \(P_2\) and not \(P_1\).
Some may argue that truth conditions and structure are not the basic concepts for a theory of meaning. Rather, a theory of meaning should take assertion conditions and structure to be the cognate basic concepts, and truth conditions enter the theory as a theoretical concept\textsuperscript{12}. I do not wish to enter into that dispute, and concede that a theory of meaning as described above may need underpinning by a theory employing more basic concepts.

Taking the language users to realise a psychological functional theory, we attribute to an utterer not only an actual set of psychological states, but also other possible states in other possible but counterfactual circumstances. And we wish the theory of meaning to reflect this feature of the functional theory. To do so, should we take the possible worlds over which the truth conditions are defined to be Kripkean, or should the domains of the various possible worlds be disjoint, with counterpart relations between their respective denizens? If we choose the latter, then the sentences "Cicero is Cicero" and "Cicero is Tully" differ in truth conditions, the first being necessary and the second contingent, and they therefore differ in structure - since same structure, for context free sentences, implies same truth conditions. On the other hand, if we choose the former, then "Cicero is Cicero" and "Cicero is Tully" have the same truth conditions, both being necessary, but, since competent language users can believe that Cicero is Cicero and not believe that Cicero is Tully, they have different syntactic structures.
I find Kripke's arguments for trans-world identity rather than counterpart relations persuasive, so the range of possible counterfactual circumstances is measured, so far as the functional psychological theory is concerned, by Kripkean possible worlds.

It may be attractive, nonetheless, to think of the truth conditions of "Cicero is Tully" being measured by non-Kripkean, counterpart-related possible worlds, if the members of the language community uniformly associate one concept or recognitional ability with "Cicero" and another concept or recognitional ability with "Tully". For then we could say, even though it is metaphysically impossible that Cicero is not Tully, nonetheless, the members of this language community can conceive of Cicero not being Tully. And then one could argue that the theory of utterance meaning which most closely aligns the contents of utterance with belief expressed, so as to validate the assertion-belief generalisation, that theory of meaning should define truth conditions across such counterpart-related possible worlds.

However, we should reject this line of argument. For whilst some competent language users may associate one concept with "Cicero" and another with "Tully", we have no reason to suppose that those who know that Cicero is Tully do so. The classicist will not be able to divide whatever individuating information he possesses between the names, will not be able to conceive of Cicero not being Tully. And we want the classicist to speak the same
language, so far as the names "Cicero" and "Tully" are concerned, as the relative ignoramus. So let the truth conditions be Kripkean, and the ignorant can be accommodated by assigning different structures to different proper names.⁷³

Let us note that this decision implies that there will be cases where two sentences $P_1$ and $P_2$ differ in meaning, because they have different structures, although they have the same truth conditions, and there is no a priori route from $P_1$ to $P_2$, or vice versa. The sentences "Cicero is Cicero" and "Cicero is Tully" are an example.

Further, whether two words or phrases have the same syntactic structure or not is a theoretical issue, not merely a matter of the letters on the page. To adapt some examples of Tyler Burge's, the following pairs of sentences have the same syntactic structures:

Paul returns in a fortnight ............. (1)
Paul returns in fourteen days ............ (2)

as have

Paul has arthritis .......................... (3)
Paul has a rheumatic disease of the joints (4)

For (1) and (2) have the same truth conditions, as have (3) and (4). And, we hold, no-one can believe that (1) without believing that (2), and vice versa. So they have the same structure. And (3) and (4) likewise.

So explicit definitions like

\[ \text{fortnight} =_{df} \text{period of fourteen days} \]
\[ \text{arthritis} =_{df} \text{rheumatic disease of the joints} \]

record that the seemingly syntactically simple units
"fortnight" and "arthritis" have, from the point of view of our theory of meaning, the syntactically complex structures "period of fourteen days" and "rheumatic disease of the joints", respectively. Whence one who sincerely utters both

Paul will return in a fortnight ........ (5)
and Paul will return in ten days ........ (6)
does not believe that (5) and believe that (6), no-one can. Rather, he is not in this respect a competent language user. Perhaps he speaks an ideolcet in which "fortnight" means a period of ten days. The assertion-belief generalisation does not apply to these utterances of his.

So our decision that "Cicero", "Tully", etc. have different syntactic structures was, in effect, a decision to count sincere utterers of "Cicero is Tully" and utterers of "Cicero is not Tullv" as utterers of the same language. We thereby recognise that it is possible to believe that Cicero is Cicero and believe that Cicero is not Tully. We recognise the possibility that everyone might believe, say, that Jack the Ripper was the Duke of Clarence, and their beliefs be false. By contrast, everyone couldn't believe that a fortnight is ten days and be wrong. If everyone sincerely, reflectively uttered "A fortnight is ten days", then the belief so expressed would be true, "fortnight" would mean a period of ten days.

Nor is our decision merely arbitrary. If everyone comes to assert "Jack the Ripper was the Duke of Clarence", 
we cannot make their assertion true by a change of ideoclect. We might find a single composite entity to be the common referent of both "Jack the Ripper" and "the Duke of Clarence", but other beliefs would be made massively false—e.g. that Jack the Duke has the same number of bodies as the rest of us.

The psychological theory outlined in the last section meshes with this theory of meaning, so far as proper names are concerned. We do not have

\[ a = b, \quad z \text{ bel } Fa \vdash z \text{ bel } Fb \]
because we do not have \( Fa \vdash Fb \), as rule (C) would require. Further, the rules \( \vdash \) are revealed as rules of syntactic transformation such that \( P_1 \vdash P_2 \) if and only if \( S_{P_1} = S_{P_2} \). For if \( P_1 \vdash P_2 \) then \( T_{C_{P_1}} = T_{C_{P_2}} \) and, by rule (C), it is impossible to believe that \( P_1 \) and not believe \( P_2 \) and vice versa. So the psychological theory tells us that, in general, where \( P_1 \vdash P_2 \), i.e. where there is an \( a \) priori route from \( P_1 \) to \( P_2 \) and vice versa, \( P_1 \) and \( P_2 \) have the same syntactic structure. I conjecture that the reverse implication also holds: if we do not have \( P_1 \vdash P_2 \), then it is possible for someone to believe that \( P_1 \) and not \( P_2 \), or vice versa, even if \( T_{C_{P_1}} = T_{C_{P_2}} \), and hence \( S_{P_1} \# S_{P_2} \).

A case in point would be Kripke's Pierre. So understood, \( \vdash \) are rules of syntactic transformation such that \( P_1 \vdash P_2 \) is both sufficient and necessary for \( P_1 \) and \( P_2 \) having the same syntactic structure.

Thus the theory of meaning supports the view taken in Appendix B to Chapter Four, that a functionalist should
not attempt to incorporate syntactically inconsistent beliefs into his functional theory. Such beliefs are impossible coexistent functional states. Which is not to say that a functionalist need deny that we do sometimes hold such beliefs: but we do so by realising two functional theories, there being no one global functional explanation of the totality of our behaviour.

Entities which realise the same functional theory may be in different functional states. No surprise, then, that psychological theory seems to recognise as alternative possible states:

\[ z \text{ bel of } \langle a_1 b_2 \rangle \text{ that } [a_1 = b_2] \text{ .......... (7) } \]
\[ z \text{ bel of } \langle a_1 b_2 \rangle \text{ that } [a_1 \neq b_2] \text{ .......... (8) } \]

In general the psychological theory permits all combinations of belief about the identities and differences of the bearers of proper names.

One psychological subject, because of the identifying information or recognitional abilities he associates with each name, may be able to conceive of a being b and of a not being b. (Perhaps he is a detective on the track of the notorious Mr Hyde, and can contemplate the epistemic possibility that Mr Hyde is Dr Jeckyl, and also contemplate the epistemic possibility of Mr Hyde not being Dr Jeckyl. As a student of Kripke, our detective knows that one of these suppositions is metaphysically impossible, but not, of course, which.) However, another psychological subject, given the identifying information and recognitional abilities he associates with the names cannot conceive of
a\neq b, if he believes \( a=b \). And a third subject, who believes \( a\neq b \), cannot conceive of \( a=b \).

For the first subject (7) and (8) do report alternative functional states, but for the second subject, (7) characterises his functional system. And for the third (8) characterises his functional system. For the first (7) and (8) are alternative 'inputs'; for the second and third, (7) and (8) respectively characterise their 'machine tables'.

The three psychological subjects realise a common generic functional system, but realise different specific variants on the theme.

In general, I conjecture, where \( P_1 \) and \( P_2 \) have the same truth conditions but different structures, i.e. there is no a priori route from \( P_1 \) to \( P_2 \) and vice versa, so

\[
\begin{align*}
& z \text{ bel} [P_1] \quad \ldots \ldots \quad (9) \\
& z \text{ bel} [P_2] \quad \ldots \ldots \quad (10)
\end{align*}
\]

are mutually independent, then if (9) or (10) is true of a subject, it may mark one of a pair of possible states of that subject ((10) or (9) being the other), or it may characterise the functional system of that subject. In which last case, the other report does not report one of a pair of possible psychological states of that subject.

Thus what appears to be a common psychological theory of the language users may really be the description of a family of different functional theories. Only in the limiting case, where a subject has the maximal array of possible functional states, is the common psychological theory the theory of that subject.
Nonetheless, all the subjects, whatever the differences between their specific functional systems, grasp and understand the same language. For all assign different meanings to $P_1$ and $P_2$. All recognise that someone, perhaps not themselves, can believe that $[P_1]$ and not that $[P_2]$, or vice versa, since all recognise that there is no a priori route from one to the other. Thus the common functional theory which each language user applies to himself and to others has a greater multiplicity of functionally possible states than may actually be the case for any of the individual language users.

Hence, when an explanatory complex contains a clause of the form

$$z \text{ bel of } \langle \ldots x_1 x_2 \ldots \rangle \text{ that } [\ldots & x_1 = x_2 & \ldots]$$

or a clause of the form

$$z \text{ bel of } \langle \ldots x_1 x_2 \ldots \rangle \text{ that } [\ldots & x_1 \neq x_2 & \ldots]$$

that clause may record an epistemically contingent functional state of $z$, or it may record a fixed feature of $z$'s current functional system.

So far our language has contained proper names but no indexical elements, and so our theory of meaning has not needed to distinguish sentence types from token utterances. Now let us suppose that the language has indexical terms. No doubt a full theory of meaning for such a language would specify, for each indexical term, a function from the context of utterance of the indexical to the value. Thus, in the case, for example, of "I" the function goes from the utterance to the utterer; for the past tense from the time of utterance to all earlier times, etc. And no doubt such a
theory of meaning should explain the a priori inferential relations between tokens of, e.g., "It will rain", "It is raining" and "It did rain" said on successive days.\textsuperscript{15}

However, I propose that our theory of utterance meaning simply assigns to each token indexical term its semantic value, ex cathedra, and records the structure of each token singular indexical term occurring in a token utterance by subscripting a numeral. Two token indexical terms have the same structure and receive the same numerical subscript if and only if it is impossible to believe that they have different values. If it is possible for some competent language user to believe the two token indexicals have different values, then they receive different numerical subscripts.

Of course, this process of marking the same or different structures of token indexical terms by same or different numerals, is precisely the process of attaching numerical subscripts to the indexicals in paratactic clauses when assembling various attitude reports into an explanatory complex. When subscripting indexicals within each paratactic clause different subscripts were given if and only if it was possible to believe that they had different values. However, the numerals subscripted in each clause were disjoint, but with the sole exception that all the numerals were repeated in conjuncts of the forms \( x_i = x_j \) and \( x_k \neq x_L \) in the belief clause. We have seen that these clauses play a special role in psychological explanations: they link up and mutually orient the various clauses into
an explanatory complex. And we have seen also that \( x_i = x_j \) and \( x_{i_k} \neq x_{j_k} \) may either characterise alternative states of a subject's functional system, or may characterise that system.

Section (iv): Quantification-in and the truth definition

Quantifiers binding and quasi-binding are kosher only if we can give an extensional account of an arbitrary sequence \( s \) satisfying such a sentence of the form

\[
(\exists x)(\ldots\langle\ldots x \ldots\rangle\ldots [\ldots x \ldots])\ldots
\]

an account which will fit an otherwise standard definition of truth.

Let us first see what the problems are. Take the standard account of quantification

\[
s \text{ sat } (\exists x_i)F_{x_i} \iff (\exists s')(j)((i \neq j \rightarrow M_{s'j} = M_{s'j}) \& s' \text{ sat } F_{x_i})
\]

where "Msj" denotes the \( j^{th} \) member of \( s \), and \( s' \) sat \( F_{x_i} \) if and only if the \( i^{th} \) member of \( s' \) satisfies \( F \). On this view

\[
s \text{ sat } (\exists x_i)(z \text{ bel of } \langle\ldots x_i\ldots\rangle \text{ that}[F\ldots x_i\ldots])
\]

&

\[
z \text{ des of } \langle\ldots x_i\ldots\rangle \text{ that}[G\ldots x_i\ldots])\ldots(1)
\]

for example, if and only if there is some object, say \( a \), such that

\[
z \text{ bel of } \langle\ldots a_i\ldots\rangle \text{ that}[F\ldots x_i\ldots] \& z \text{ des of } \langle\ldots a_i\ldots\rangle \text{ that}[G\ldots x_i\ldots] \]

\[
.............(2)
\]

where \( a \) is given as the value of the indexicals "\( x_i \)" in the two paratactic clauses. But (2) contains no information as to whether \( z \) takes this object of his belief to be the same or different from that object of his desire, nor whether, if the object \( a \) turns out to be the value of some
other indexical in \([F \ldots x_1 \ldots]\) or in \([G \ldots x_1 \ldots]\), then \(z\) believes their values to be the same or not. The subscripts in the \([\ldots]\) clauses merely relate each indexical to a place in the associated \(<\ldots>\) at which its value is to be found. So, if (2) is equivalent to (1), (1) cannot contain this information either, and we are back to merely binding the variables in \(<\ldots>\), not also thereby quasi-binding the variables in \([\ldots]\).

We cannot solve this problem by saying that (1) is equivalent to there being an object, say \(a\), such that

\[
\text{z bel of } \langle \ldots a_1 \ldots \rangle \text{ that } [F \ldots a_1 \ldots] \quad \& \quad \text{z des of } \langle \ldots a_1 \ldots \rangle \text{ that } [G \ldots a_1 \ldots]
\]

We would then bind, not quasi-bind, the variables, inducing referential transparency. For example, suppose we have

\[
(\exists x_1 x_2) (z \text{ bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 x_2])
\]

and suppose, in fact, the relevant \(x_1=x_2=a\). Then the above has the same truth value as

\[
z \text{ bel of } \langle a_1 a_2 \rangle \text{ that } [Fa_1 a_2]
\]

whether or not \(z\) realises the two res are the same.

Perhaps we should try another tack. Given an explanatory complex, we want to define

\[
s \text{ sat } (\exists x_1 \ldots x_n)(\ldots\ldots\ldots)
\]

such that we can eliminate the quantifiers to reveal the satisfaction conditions of the resulting \((\ldots\ldots\ldots)\) with values assigned the variables \(x_1, \ldots, x_n\). Suppose we proceed by retaining all the subscripts in \((\ldots\ldots\ldots)\), and by allowing them to symbolise the subject's views as to
the identities and differences of the various res, as they
did in the original explanatory complex. For example

\[(\exists x_1 x_2)(z \text{ bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 x_2] \& z \text{ des of } \langle x_1 x_2 \rangle \text{ that } [Gx_1 x_2])\]
is true if and only if there are entities, a and b, say,
which may or may not be the same, such that

\[z \text{ bel of } \langle a_1 b_2 \rangle \text{ that } [Fx_1 x_2] \& z \text{ des of } \langle a_1 b_2 \rangle \text{ that } [Gx_1 x_2]\]

where, now, we regard the subscripts as not only indicating
the place in \langle \rangle at which the value of the given indexical
is to be found, but also z's views as to the identities
and differences between the res of his two attitudes.
Alas, the truth definition is no longer extensional. We
have now a sentence of the form A&B, but its truth depends
upon more than the truth of A and the truth of B. It
depends also upon how z takes the attitudes reported by
each conjunct to be related.

Reviewing these problems brings us close to their
solution. Let us take, as a test case, an explanatory
complex:-

\[(\exists x_1 x_2 x_3 x_4)(z \text{ bel of } \langle x_1 x_2 x_3 x_4 \rangle \text{ that } [Fx_1 x_2 x_3 x_4] \& z \text{ des of } \langle x_1 x_2 \rangle \text{ that } [Gx_1 x_2] \& z \text{ int of } \langle x_3 x_4 \rangle \text{ that } [Hx_3 x_4])\] \hspace{1cm} (1)

Now let us rewrite this in the style:-

\[(\exists x_1 x_2 x_3 x_4)(z \text{ bel of } \langle x_1 x_2 x_3 x_4 \rangle \text{ that }^1 \& z \text{ des of } \langle x_1 x_2 \rangle \text{ that }^2 \& z \text{ int of } \langle x_3 x_4 \rangle \text{ that }^3)\]

\[\text{that }^1 = Fx_1 x_2 x_3 x_4\]
\[\text{that }^2 = Gx_1 x_2\]
\[\text{that }^3 = Hx_3 x_4\] \hspace{1cm} (1')
All that has happened is that we have separated out the paratactic clauses, using superscripts 1 to 3 to relate each demonstrative to its appropriate paratactic clause. As before the subscripts in the paratactic clauses are indexicals whose values are the same-subscripted members of the relevant sequence. And as before, the same subscript for two indexicals indicates that z takes the attitudes so reported to have the same res. So far, no substantial change.

If a, b, c, d are the entities which make (1) true, where they may or may not be distinct entities, we want (1) to be true because

\[
\begin{align*}
&\text{z bel of } \langle a_1 b_2 c_3 d_4 \rangle \text{ that}^1 \& \text{z des of } \langle a_1 b_2 \rangle \text{ that}^2 \\
&\quad \& \text{z int of } \langle c_2 d_4 \rangle \text{ that}^3 \& \text{zSIV(that}^1,\text{that}^2,\text{that}^3) \\
&\text{that}^1 = Fx_1 x_2 x_3 x_4 \\
&\text{that}^2 = Gx_1 x_2 \\
&\text{that}^3 = Hx_3 x_4
\end{align*}
\]

(2)

where, to put it very roughly, "zSIV(\text{that}^1,\text{that}^2,\text{that}^3)" says that z takes indexicals with the same subscripts in that^1, that^2 and that^3 to have the same value.

We need to refine this idea, because that^1-3 are utterances of the attitude reporter, not z, and z may be without access to that^1-3. We met a similar problem when trying to formulate an assertion-belief generalisation in Chapter Two. Then we said that the indexicals in the token t (to which the believer is disposed to assent) and in the token p (produced by the belief reporter) may differ, to allow for the different contexts of believer
and belief-reporter, provided (i) that they retain the same semantic values, and (ii) where a competent language user could take two indexicals in \( t \) to have different values, the corresponding indexicals in \( p \) must be such that a competent language user could take them to have different values. We need to extend these rather hazy ideas to attitudes other than belief.

Let us suppose that for each attitude involved in an explanatory complex (belief, perception, desire, intention) there is a token to which \( z \) is disposed to a sincere appropriate performative response: if only saying to himself, for example, "Yes, that is what I want". And we require (i)indexicals in the demonstrated clauses that have the same values as in these tokens, and (ii)where a competent language user could take the indexicals in those tokens to have different values, the demonstrated clauses employ different indexicals. Now we can say:

\[
SIV(z's \text{ bel that}^1, z's \text{ des that}^2, z's \text{ int that}^3) = df
\]

\( z' \)s belief reported by that\(^1\) and \( z' \)s desire reported by that\(^2\) and \( z' \)s intention reported by that\(^3\) are such that where that\(^1\)-\(^3\) employ the same indexical \( z \) takes the relevant indexicals in the relevant tokens to have the same values - i.e. where that\(^1\)-\(^3\) employ the same indexical, \( z \) takes the relevant ref of his relevant attitudes to be identical.

And which of \( z' \)s beliefs does "\( z' \)s bel that\(^1\)" denote? The one reported elsewhere in the same sentence by also referring to that\(^1\), of course. And "\( z' \)s des that\(^2\)",
"z's int that\textsuperscript{3}" likewise. Let us abbreviate

\[ \text{SIV}(z's \text{ bel that}\textsuperscript{1}, z's \text{ des that}\textsuperscript{2}, z's \text{ int that}\textsuperscript{3}) \]

so understood, to

\[ \text{zSIV}(\text{that}\textsuperscript{1}, \text{that}\textsuperscript{2}, \text{that}\textsuperscript{3}). \]

Now we are almost home and dry. We want (1) to be true because (2) is true. But we have jumped straight from (1) which is prefixed by four quantifiers, to (2), which is bereft of quantifiers. We need a general rule by which we may strip off just one quantifier at a time.

Suppose

\[ (\exists x_1)(F^j \ldots \langle \ldots x_i \ldots \rangle \ldots [\ldots x_1 \ldots]) \]

is an explanatory complex, where (\exists x_1) is the leading quantifier of widest scope, and there are \( j \) clauses demonstrated in the course of the explanatory complex. (Of course "\( x_1 \)" may recur in more than one sequence, and in more than one demonstrated clause.) We want

\[ s \text{ sat } (\exists x_1)(F^j \ldots \langle \ldots x_i \ldots \rangle \ldots [\ldots x_1 \ldots]) \]

if and only if

\[ (\exists s')(n)(n \neq i \rightarrow Msn = Ms'n & s' \text{ sat } (F^j \ldots \langle \ldots x_i \ldots \rangle \ldots [\ldots x_1 \ldots]) \]

\[ & \text{zSIV}(x_1, \text{that}\textsuperscript{1}, \ldots, \text{that}\textsuperscript{j}) \]

where, of course, "\( \text{zSIV}(x_1, \text{that}\textsuperscript{1}, \ldots, \text{that}\textsuperscript{j})" says that where "\( x_1 \)" recurs in any of that\textsuperscript{1}, \ldots, that\textsuperscript{j}, \( z \) takes the relevant rest of his relevant attitudes to be identical. And

\[ s' \text{ sat } (F^j \ldots \langle \ldots x_i \ldots \rangle \ldots [\ldots x_1 \ldots]) \]

\[ & \text{zSIV}(x_1, \text{that}\textsuperscript{1}, \ldots, \text{that}\textsuperscript{j}) \]

if and only if there is some entity, say \( a \), such that

\[ (F^j \ldots \langle \ldots a_i \ldots \rangle \ldots [\ldots x_1 \ldots]) \& \text{zSIV}(x_1, \text{that}\textsuperscript{1}, \ldots, \text{that}\textsuperscript{j}) \]
And now we have it. Successive elimination of the quantifiers \((\exists x_1 \ldots x_n)\) from an explanatory complex will successively eliminate the variables from the \(\langle \quad \rangle\) terms, and introduce a new conjunct: first "\(\text{zSIV}(x_1, \text{that}^1, \ldots, \text{that}^j)\)". then "\(\text{zSIV}(x_1 x_2, \text{that}^1, \ldots, \text{that}^j)\)". then ... , and finally "\(\text{zSIV}(x_1 \ldots x_n, \text{that}^1, \ldots, \text{that}^j)\)".

Thus the role which the quantifiers performed – linking up and mutually orienting the various content clauses – has been taken over by the new conjunct. And in such a way that we can incorporate the quantification and quasi-binding of variables in paratactic clauses into an extensional truth definition.

Such quantification and quasi-binding to form explanatory complexes only applies to language using creatures; creatures who can refer to the various res of their attitudes, and who can take the referring expressions to have, or not to have, the same values. And obviously the theory only applies to the behaviour which dares to speak its name – sometimes we would not acknowledge our motives even to ourselves. But I hope that the cases to which the theory does apply are the central cases; the Old Adam of a spreading family.

Section (v): Hallucination

I argued in Chapter Four that narrow functionalism does not give a plausible account of the subject's grasp of existential quantification. Narrow functionalism cannot distinguish a reading of "\(\exists\)" as "there is" and
reading it as "there may be", with variable modality. And narrow functionalism cannot explain why classical inference rules preserve truth (Myself having a winning strategy), nor why inductive strategies preserve the appropriate semantic values. In this chapter I have argued that explanatory complexes require external quantifiers to bind and quasi-bind variables.

My project seems threatened by honest-to-God hallucination, cases in which the subject hallucinates some apparent res, say hallucinates a dagger before him, not merely mistakes some real shadown for a dagger. For it is plausible to claim that (i) Tom hallucinating a dagger is in the same psychological state as his narrow doppelganger Twin Tom who is actually perceiving a dagger, and that (ii) the same psychological explanation applies to Tom the dagger-hallucinator as to his twin, the dagger-perceiver. But, of course, I cannot agree with these claims: the explanatory complex which applies to Twin Tom starts with a prefix "(∃x₁...xₙ)", whilst Tom's, if it is to be true, can only start with "(∃x₁...xₙ₋₁)", there being no res he thinks is a dagger. So it behoves me to say what psychological state Tom the dagger hallucinator is in, and what psychological explanation does apply to his cogitations. And that explanation had better make plausible why it seems to Tom the hallucinator just as it seems to his perceiving twin.

Provided the above can be achieved, hallucinations do not threaten wide functionalism. The empirical datum
is that we do sometimes hallucinate, not that we hallucinate all the time, being brains in vats. Wide functionalism can allow for hallucinations, provided enough of our interactions with the world are not hallucinatory for us to realise the requisite wide functional system. Wide functionalism is even consistent with all the apparent res of our attitudes being hallucinatory for certain restricted periods - as when we dream.\textsuperscript{16}

Indeed, naturalism may require that any given putative res may be an hallucination - but not all res all the time. For naturalism sees us as creatures on an ontological par with the objects in our environment, and only contingently causally connected with them. (In contrast to Idealism: Kantian or Berkelean.) Our thinking that there is a so-and-so cannot make it so: it can seem to us just as if there is whilst, in fact, there is not. So a functionalism, wide or narrow, which is motivated by naturalism, should provide for the possibility of any apparent res being an hallucination.

However, a naturalism which holds that psychological facts are determined by wide physical functional facts can also respect the intuition that, roughly, one cannot be a dagger hallucinator unless one is also, sometimes, a dagger perceiver. This dictum needs some fine-tuning, since one can hallucinate a unicorn, never having been a unicorn perceiver. But then one must have perceived white things and horses, or perhaps just pictures of horses or unicorns, or ... providing some basic components for one's conception
of a unicorn. Further fine-tuning may be required. Pigeon's eyes, I believe, are responsive to wavelengths to which our eyes are not. I can well believe that, consequentially, pigeons can see colours which we cannot, and cannot even imagine. But it may be that, tripping on LSD, I hallucinate pigeon colours, colours I have never perceived, and am not then really perceiving under the influence of the drug. But, even if so, I could not report the specific hallucinations, my language having no specific colour concepts appropriate to those hallucinations. So it may be that the specific reports of hallucinations, and psychological explanations employing them, require perceived instances of basic components.

Without further ado, what psychological state is Tom the dagger-hallucinator in, and is Twin Tom the dagger seer in, such that it seems to Tom just as it seems to Twin Tom?

Of Twin Tom, we have
\[(\exists x)(T-x \text{ per of } \langle x_1 \rangle \text{ that } [Dx_1])\]
Tom is clearly in the state
\[-(\exists x)(T-x \text{ per of } \langle x_1 \rangle \text{ that } [Dx_1]) \& T-x \text{ per } [(\exists x)(Dx)]\]
and this attributes different psychological states to Tom and Twin Tom. But, so far, we have not shown why it seems to Tom as it seems to Twin Tom.

Let us introduce the function \(\text{PRPIS}(T-x, w)\), whose value, for Tom in possible world \(w\), is the psychologically relevant physical inner state of Tom in \(w\). If Tom is in some wide psychological functional states in \(w\), then those
states are realised by some set of wide physical states in \( w \), and \( \text{PRPIS}(\text{Tom},w) \) is simply that proper part of those physical states which is bounded by Tom's bodily envelope. We can now say, of hallucinating Tom, that there is some possible world, in which his psychologically relevant physical inner states are the same, but in which he is perceiving a dagger. I.e.:

\[
\{ \exists w \} \{ (\exists x) (\text{Tom per of } <x_1> \text{ that } [\text{D}_x \text{.}]) \ & \& \text{PRPIS}(\text{Tom},w) = \text{PRPIS}(\text{Tom},a) \}
\]

where \( a \) is the actual world.

Note that we do not expect that Tom's total internal physical states are the same in \( w \) as in \( a \). For his hallucination will have an internal physical cause - cheese for supper - whereas in \( w \) the experience has an external physical cause, the dagger. So although cheese eating is, in one sense, a psychologically relevant physical event, we don't want it to be part of \( \text{PRPIS}(\text{Tom},a) \). We want \( \text{PRPIS}(\text{Tom},a) \) to denote those internal physical states of Tom which are a proper part of the physical states which realise Tom's psychological states. Further, we expect \( \text{PRPIS}(\text{Tom},a) \) to be bounded by some envelope more restricted than the bodily envelope. For, no doubt, the dagger seer and the dagger hallucinator will differ internally in that the former has relevant retinal activity which the latter lacks. So we will need to dig deeper to find a common core of psychologically relevant neuronal activity. But I ignore such complications.

In sum, hallucinating Tom is in the following psychologically relevant state:-
Tom per \([(\exists x)(Dx)] \land -[(\exists x)(\text{Tom per } of \langle x \rangle \text{ that } [Dx_1])] \land \{w \}[(\exists x)(\text{Tom per } of \langle x \rangle \text{ that } [Dx_1]) \land \text{PRPIS}(\text{Tom}, w) = \text{PRPIS}(\text{Tom}, a)]\}

Why does it seem to hallucinating Tom just as it seems to perceiving Twin Tom? Clearly I must claim that it is (psychologically) with perceiving Twin Tom in a as it is with Tom in w. And, given PRPIS(Tom, a) = PRPIS(Tom, w), it seems to Tom in a as it seems to Tom in w. So it seems, cognitively, to Tom in a as it is with perceiving Twin Tom in a.

This is not a retreat to narrow functionalism. The physical states PRPIS(Tom, a) are not alone sufficient to sustain a psychological life, let alone that Tom and Twin Tom seem to share. PRPIS(Tom, a) is a proper part of a wide set of physical states. PRPIS(Tom, a) and PRPIS(Twin Tom, a) are also proper parts of different wide sets of physical states. Tom-in-the-vat may share PRPIS(Tom, a), but since they are not in his case a proper part of a wide functional system, he lacks a psychological life.

I want to say that where Twin Tom takes the res of one attitude to be the same as, or different from, that of another attitude, hallucinating Tom takes the apparent res of his corresponding attitudes to be the same or different. I hope, of course, that PRPIS(Tom, a) = PRPIS(Twin Tom, a) is a sufficient condition for this to be so. But, given an explanatory complex applying to perceiving Twin Tom, I have not yet given an explanatory complex applying to
hallucinating Tom, where the two explanatory complexes are different, yet they show the correspondence between their attitudes, and between their judgements of identity and difference of res, real or apparent. Let me now try to do this.

Suppose, for example, Twin Tom sees a dagger, desires to use it and intends to grasp it. Then we have an explanatory complex of the form:

\[(\exists x_1 x_2 x_3)(\langle \text{T-Tom}\rangle \text{bel of } \langle x_1 x_2 x_3 \rangle \text{that } \langle Fx_1 \& x_1 = x_2 \& x_1 = x_3 \rangle) \& \langle \text{T-Tom}\rangle \text{des of } \langle x_1 \rangle \text{that } \langle Gx_2 \rangle \& \langle \text{T-Tom}\rangle \text{int of } \langle x_3 \rangle \text{that } \langle Hx_3 \rangle \]

Tom hallucinates a dagger. But if we write something along the lines of

\[(\exists x)(Fx) \& \langle \text{T-Tom}\rangle \text{bel that } \langle (\exists x)Fx \rangle \& \langle \text{T-Tom}\rangle \text{des that } \langle (\exists x)Gx \rangle \& \langle \text{T-Tom}\rangle \text{int that } \langle (\exists x)Hx \rangle \]

we no longer have a psychological explanation, for we have lost the connection between the various attitudes.

Perhaps we can say of hallucinating Tom

\[(\exists x)(Fx) \& \langle \text{T-Tom}\rangle \text{bel that } \langle Fx_1 \& x_1 = x_2 \& x_1 = x_3 \rangle \& \langle \text{T-Tom}\rangle \text{des that } \langle Gx_2 \rangle \& \langle \text{T-Tom}\rangle \text{int that } \langle Hx_3 \rangle \& \langle \text{T-Tom}\rangle \text{SIV } (x_1 x_2 x_3, \langle 1 \rangle, \langle 2 \rangle, \langle 3 \rangle) ?

But Tom has no access to that \(1-3\). Tom is disposed to respond appropriately to tokens which do manifest his attitudes. How are we to link these to that \(1-3\)? The indexicals of those tokens lack semantic values, as do those of that \(1-3\). But, since we know that \(1-3\) lack semantic values, or stipulate that they do, we can require that the tokens that \(1-3\) have the same linguistic meaning as the tokens which Tom would appropriately respond to.
Call this new relation

\[ \text{T} \text{om SIV}'(x_1x_2x_3, \text{that}^1, \text{that}^2, \text{that}^3) \]

The explanation of dagger-seeing \( \text{T} \text{om}'s \) behaviour being (1), the explanation of dagger-hallucinating \( \text{T} \text{om}'s \) behaviour is:

\[-(\exists x_1x_2x_3)(\text{Tom bel of } \langle x_1x_2x_3 \rangle \text{that}^1[Fx_1 \& x_1 = x_2 \& x_1 = x_3] \]
\& \text{Tom des of } \langle x_2 \rangle \text{that}^2[Gx_2] \& \text{Tom int of } \langle x_3 \rangle \text{that}^3[Hx_3] \]
\& \text{Tom bel that}^4[Fx_1 \& x_1 = x_2 \& x_1 = x_3] \& \text{Tom des that}^5[Gx_3] \]
\& \text{Tom int that}^6[Hx_3] \]
\& \text{Tom SIV}'(x_1x_2x_3, \text{that}^4, \text{that}^5, \text{that}^6) \]

(2)

We cannot expect, in general, the demonstrated clauses that \( ^4-6 \) of (2) to be syntactically the same as that \( ^1-3 \) of (1). For that \( ^1-3 \) of (1) may have required changes of indexicals to achieve the same semantic values as the tokens Twin Tom would respond to. Whereas that \( ^4-6 \) have the same linguistic meanings as those tokens which Tom would respond to.

Hopefully, given (1) and (2), we can infer that it seems to hallucinating Tom, of whom (2) is true, as it seems to his perceiving twin, of whom (1) is true. For suppose (1) is made true by some objects \( a, b, c \) which may or may not be different, such that:

\[ \text{T} \text{-Tom bel of } \langle a_1, b_2, c_3 \rangle \text{that}^1[Fx_1 \& x_1 = x_2 \& x_1 = x_3] \]
\& \text{T} \text{-Tom des of } \langle b_2 \rangle \text{that}^2[Gx_2] \& \text{T} \text{-Tom int of } \langle c_3 \rangle \text{that}^3[Gx_3] \]
\& \text{T} \text{-Tom SIV}(x_1x_2x_3, \text{that}^1, \text{that}^2, \text{that}^3) \]

(3)

The tokens which Tom is disposed to respond appropriately to, and which warrant the conjuncts of (3), are of
the same type as the tokens which Tom is disposed to respond to in parallel fashion, and which warrant the corresponding conjuncts of (2). And the behavioural warrant for "T-Tom SIV(\(x_1x_2x_3\), that\(^1\),that\(^2\),that\(^3\))" is of the same type as the behavioural warrant for "Tom SIV'(\(x_1x_2x_3\),that\(^4\),that\(^5\),that\(^6\)).

Finally, we need to generalise to mixed-mode cases. A Tom may hallucinate some but not all of the res which his twin perceives, and which are involved in the same complex explaining that twin's behaviour. Clearly, we may now expect there to be tokens to which Tom is disposed to respond appropriately, and corresponding 'that' clauses in the complex explaining his behaviour, where some of the indexicals have semantic values, and some do not. When we move from Tom's context to the attitude reporter's context we wish to preserve information on what Tom takes to be the identities and differences between the various res, real and apparent. We do this by a composite rule: where an indexical has a semantic value, we change it to preserve that value, if necessary (subject to the restrictions discussed in section (iv)); and where the indexical has no semantic value, we leave its linguistic meaning unchanged. There is no danger that an indexical originally empty will achieve a semantic value in the new context of the attitude belief-reporter. For the can explicitly deny the existence of such a res, as illustrated in (2) above. We need a new polyadic relational term "zSIV'"(\(x_1\ldots x_n\),that\(^1\),...,that\(^j\))" to record this mixed-mode relation between \(x_1\ldots x_n\) in that\(^1\)\ldots \(^j\) and the indexicals in the tokens to which z responds appropriately.
Section (vi); Peacocke on modes of presentation

Peacocke develops an account of demonstrative modes of perceptual presentation which will serve to link up and mutually orient concurrent attitudes, without a need to quantify into them. But, as we shall see, even if we suppose that such modes of presentation are sufficient for the mutual rationality of synchronic attitudes, they do not explain diachronic rationality: demonstrative modes of presentation cannot link up and mutually orient temporally separated attitudes.

Peacocke claims:

no set of attitudes gives a propositional attitude explanation of a person's acting on a given object unless the content of those attitudes includes a demonstrative m.p. of that object.

Demonstrative m.p.'s are distinguished by type by their various constitutive roles:

what is distinctive of a given demonstrative type is the pattern of evidence or prior states to the holding of which judgements containing tokens of that type must be sensitive; and the constitutive role associated with the given type is intended to capture this complex pattern of evidential sensitivity

For example, the demonstrative mode of presentation [self] has its constitutive role specified by: the person who has these experiences. And the demonstrative mode of presentation [now] by: the time this attitude (thinking, belief, intention, etc.) occurs. And [that bowl] by: the bowl responsible, in the way required for perception, for the experience as of a bowl in that region of my visual field - assuming it is a visually presented bowl.

In general, "the demonstratives in [specifications of]
constitutive roles are all of a special kind: they are all demonstratives referring to the thinker's current conscious states - experiences and thoughts".  

However, Peacocke distinguishes the constitutive role from the content of a thought employing a demonstrative mode of presentation which has that constitutive role: a person may have thoughts employing such modes of presentation without having the intellectual sophistication to think about his own experiences. 

So much for types of demonstrative modes of presentation. A token demonstrative mode of presentation of a given type occurs when a particular object is presented on a particular occasion by that type. 

If $\Delta$ is a type mp, then $[\Delta x]$ will be the token mp obtained by indexing the type $\Delta$ with the object $x$. 

$'[\Delta x]'$ here is a complex term, consisting of the application of a functor to the variables '$\Delta$' and '$x$'

So if Tom hallucinates a bowl, he entertains a type demonstrative mp, but not a token mp. 

However, because Peacocke is constructing a theory of Fregean, inter-subjective thoughts, the same token demonstrative mp can occur in many thoughts: for example, Tom will have many thoughts over the years employing the same token mp [self Tom]. Modes of presentation are constituents of Fregean thoughts, where many thinkers can think the same thought, as can one thinker at various times. 

The thoughts are not the individual representational states of the various thinkers of that one thought. Thoughts are not, therefore causally active in the production
of behaviour. (Some thoughts can only be entertained by one thinker: only Peter can entertain thoughts containing the constituent [self Peter]. And there are other restrictions: only in 1900 could one think a thought containing the constituent [now 1900].) \(^{27}\)

It might now seem that an explanation of why an agent acted on a given object cannot involve demonstrative mp of that object, contrary to Peacocke's claim quoted at the beginning of this section. For it is an essential property of the thought that \(P[\Delta a]\), that it is about the object \(a\), where \(a\) may be a time, place, person, or perceived object. In contrast, a given representational state causing an action is only contingently a representation of an object \(a\), rather than some other or no object at all; so we may ask for an explanation of why the agent acts on the object \(a\), rather than some other or no object. But Peacocke's conception of a thought leaves no logical room for asking that question of the thought, since that thought is essentially about that object.

However, I take Peacocke's point to be that explaining why the agent acted on a particular object \(a\) involves a demonstrative mp of \(a\) because the constitutive role of that token demonstrative mp makes the contingent connection between the particular representational state of the thinker and the object \(a\). \(^{28}\)

The special feature of demonstrative mp's is this: when you think of an object demonstratively, there is a relation in which you stand to that object and in which you stand in virtue of the relations of your current psychological states, the psychological states
Systematic variation in the subject's actions dependent upon his current psychological states will systematically alter his relations to objects demonstratively presented to him. It will not in any systematic and regular way alter his relations to objects not demonstratively presented to him. So, when Peacocke says that a set of attitudes explains why an agent acts on a given object only if those attitudes include a demonstrative mp of that object, it is the constitutive-role-plus-context which contingently determines the object - thereby determining which Fregean thought is entertained.

Peacocke's demonstrative mp's, with their constitutive roles involving the current conscious states of the agent, are well able to link up and mutually orient the agent's current attitudes at the time of action; well able, that is, to manifest synchronic rationality. If an agent has a perception whose content involves a demonstrative mp of type $A$, and a belief also involving a dem. mp of that type, ditto a desire and an intention, then he knows a priori (whether or not he can formulate the constitutive role of that type) that his perception, belief, desire and intention have the same object. Since the four attitudes are concurrent, the constitutive role in each case involves the same token aspect of the same token conscious state. Since the constitutive role is of the same type, the agent cannot help but consider the four thoughts to involve demonstrative presentations of the same object.

Hence an explanatory complex can take the form:-
z per at t that \([F[a]]\) & z bel at t that \([G[A]]\)
& z des at t that \([H[A]]\) & z int at t that \([I[A]]\)
and no quantification is needed to establish that z takes
these attitudes to have the same object.

But we shall need other techniques, 'bare' judgements
of identity and difference, I contend, to deal with
diachronic rationality. The explanation of the rational
development of Tom's detecting beliefs, which culminate
eventually in his belief that there is a trail of footprints
from library to scullery, may well involve beliefs about
individual footprints \(a,b,c,...\) demonstratively presented
at various times, such that:-

Tom bel at \(t_1\) that \([F[a]]\) & Tom bel at \(t_2\) that \([F[ab]]\)
& Tom bel at \(t_3\) that \([F[ac]]\)

Now, given \(a=b=c\), we can infer

Tom bel at \(t_1\) that \([F[a]]\) & Tom bel at \(t_2\) that \([F[a]]\)
& Tom bel at \(t_3\) that \([F[a]]\)

since the places "\(A\)" and "\(x\)" in "\([Ax]\)" are transparent. 39

But this does not tell us whether or not Tom himself
believes that he has encountered one, two or three
footprints. And the reason is, of course, that the
constitutive role of the type \(\Delta\) involves the believer's
current perceptual experiences. So the token constitutive
roles of the successive token demonstrative mp's involve
different perceptual experiences at \(t_1, t_2, t_3\), which Tom
may or may not take to be experiences of the same object.

Just as two different footprint-sightings may involve
modes of presentation of the same type, \(\Delta\), so two memory
images may be of the same type M. The constitutive role of the memory image of a footprint is: the footprint which caused this image in the way appropriate for memory. Now one memory image of a footprint sighting may be as like as we please to another, as token demonstrative modes of presentation may be of the same type. So the content of Tom's explicit or implicit judgement that these are two different token memory images (that this is the same token memory image again) of the same type, may be just his bare belief that they were caused by different footprints (by the same footprint). There may be nothing to distinguish the memory images as different token modes of presentation other than his thinking of them as memories of different footprints.

The key point is this: given that the same type of demonstrative mp is concurrently involved in more than one attitude, the thinker has an a priori guarantee that if they are token demonstrative mp's, then they are the same token demonstrative mp, presenting the same res. But, given memory modes of presentation of the same type, the thinker's judgement that they are different tokens, or the same token, need have no more content than his 'bare' judgement of identity, or difference, of objects remembered. So whilst synchronic rationality may not need quantification-in to link up and mutually orient concurrent attitudes, diachronic rationality does, for the thinker has to judge at t₂ that this footprint and that remembered footprint (sighted at t₁) are/are not the same. And, of course, that
the thinker remembers demonstrative presentations of footprints does not help: for there is no \textit{a priori} expectation that the same demonstrative type, presented at different times, is the same token.
Footnotes to Chapter Five:

1. Dennett: "Beyond Belief", in Woodfield (1)
2. Fodor J (2)
3. McGinn E: "The Structure of Content", in Woodfield (1)
4. Woodfield (1) p86/7
5. Ibid. p215
6. Ibid. p226
7. Ibid. p227
8. McGinn (2) p92
9. Fodor (2) p108
10. Perhaps descriptive modes of presentation can do the job of linking up attitudes? Perhaps each perception of a footprint involves a unique qualitative mode of presentation of the form $(\exists x)(Gx)$? We can then construct two idealised scenarios by which the detective is led rationally from individual footprint sightings to the belief that there is a trail of footprints from library to scullery:

   (A) Suppose the trail is some two hundred prints long, and the detective makes some thousand sightings in coming to his conclusion - he looks back at old prints to check alignment, to re-examine for the now obvious nick in the left heel, etc. etc. From these thousand premises with distinct qualitative contents - $F(3x)(G_1x)$, $F(3x)(G_2x)$, ..., $F(3x)(G_{1000}x)$ - he is able to conclude that there is a trail from library to scullery.

   (B) The detective employs only two indexical modes of presentation (that$_1$ and this$_2$) as follows. First the detective sees, at the same time, a pair of footprints, one below the library window and the other a little way off. So he comes to the belief that there is a trail of footprints leading from the library to that$_1$ (the second) footprint. He moves his head, keeping the that$_1$ footprint in view until he sees another footprint$_2$ on the toe-side of that$_1$ footprint. He now believes that there is a trail of footprints from the library window to this$_2$ footprint. Again, moving his head
whilst keeping this footprint in view he sees another footprint on the toe-side of this footprint, and concludes that there is a trail of footprints leading from the library window to that footprint. In this way the trail grows by successive steps, but he needs only to think, des re, of two footprints at any one time, and think of both under one or other of his two indexical modes of presentation. For he only needs to distinguish two footprints at any one time and know that the new one, because it is on the toe side of the old, is an extension of the trail as so far built up.

Perhaps Sherlock Holmes would recognise himself in (A) or (B). Perhaps his detecting is as methodical as (D), or perhaps he can think of 1000 distinct qualitative modes of presentation. Compromises between (A) and (B) are possible, but I don't recognise a realistic account of my conscious processes of reasoning in any of this. And I am not tempted by that old philosophical dodge: this is how it must be, so the quickness of the mind deceives the inner eye.

Searle (2), although he does not give de re attitudes any real role in psychological explanations, when it comes to linking up one attitude to another in explanatory complexes, he surprisingly does use variables bound by a common quantifier - p71 and p216. Of the man who says "There's a very nice sloop in the yacht harbour, and I sure wish I had it" Searle writes:--

"The mental states expressed here are first, a belief in the existence of a particular sloop and, then, a desire to have that sloop. In English,

I believe that there is a very nice sloop in the yacht harbour and I wish I had it.

Notice that in this formulation the scope of the quantifier in the content of the belief extends to the content of the desire even though the desire is not within the scope of the belief. Thus, using square
brackets for the scope of the intentional verbs and
round brackets for the quantifier and F for the intentional
content which identifies the sloop in question, we have:
\[ \text{Bel}[(\exists x)((\text{sloop } x \land Fx) \land (\forall y)(\text{sloop } y \land Fy \rightarrow y=x))] \]
& \text{Des}[I \text{ have } x] \]
This de dicto form represents the entire content of the
desire directed at a particular object."

Personally I cannot understand a sentence where the
quantifier lies within the scope of one psychological
verb but that quantifier has within its scope another
psychological verb, which second verb lies outside
the scope of the first. Searle surely owes us a
general and recursive account of the semantics of such
structures, as I try to give for my quantification-in
to bind and quasi-bind variables (in section (iv)).
Without such an account we do not know whether to take
the syntax Searle offers seriously, or to regard it as
merely ambiguous between "(\exists x)(\text{Bel}[[...x...] \land \text{Des}[[...x...]])"
and "\text{Bel}[(\exists x)(...x...)] \land \text{Des}[[(\exists x)(...x...)]]".

12 Perhaps in the manner of R Brandom, "Truth and
Assertability", J.Phil. 73 (1976)

13 It will emerge in the next chapter that my possible
worlds are only semi-Kripkean. I do not regard e.g.:
"Water is H_2O" as necessary. I shall reject the view
that "water" contains a concealed indexical. I shall
argue that "water" expresses a non-indexical conceptual
content, and the different extensions on earth and
Twin Earth are due to different communicative intentions,
which intentions will be expressed by explicit indexicals.

14 We should expect some believers to believe the
metaphysically impossible. For example, suppose that
Kripke was right when he argued that the lectern
demonstratively present could not (metaphysical
impossibility) have been made of ice. Nonetheless,
perhaps there was a student in Kripke's audience whose
attention had been wandering, but had caught something
of what Kripke was saying, and came away with the
surprising belief that that lectern (that very one there
present) was made of ice! I have no doubt that the credulous student's belief was de that very lectern. Having misheard Kripke, his attention was no doubt riveted by it; waiting to see if a puddle of water would form at its base.

If we explicate sentences containing indexicals by employing such functions, then, it seems, the functions always take widest scope. Take for example: "If I were hot, then I would leave", a counterfactual, let us suppose. We cannot explicate this as: "If the utterer of this sentence were hot, then the utterer of this sentence would leave", since the latter is true only in possible worlds in which the utterer utters that sentence, is hot, and leaves. But the former is true in possible worlds in which the actual utterer does no uttering, but is hot and leaves.

I reject out of hand Malcolm's famous view that when we dream we do not cogitate, but only dream that we cogitate.

Peacocke (1) and (2)
Peacocke (1) p205/6
Peacocke (2) p118
Ibid. p110
Ibid. p113
Ibid. p129/130
Ibid. p110
Peacocke (1) p190
Peacocke (2) p125 & p9 ft. 6
Ibid. p108
Peacocke (1) p193
Peacocke (2) p159
Peacocke (1) p190
Peacocke (2) p185
Chapter Six: Wide functionalism, concepts and properties

Section (i): Introduction

I want the wide physical functional system centred on Tom to determine the psychological facts about Tom, including, inter alia, the concepts which Tom grasps. Plausibly, the relevant physical functional facts are not all embracing, do not include a complete and most detailed history of the universe oriented upon Tom. For example, it seems clear that they do not even include the particular identities of the res of Tom's attitudes, because psychologically explanatory complexes quantify over those res, and it makes no difference to the psychological history of Tom what particular entities are the values of the variables. Thus an explanatory complex, for example

\[(\exists x_1 x_2)(\text{Tom bel of } \langle x_1 x_2 \rangle \text{ that } [Fx_1 \& x_1 = x_2] \& \text{Tom des of } \langle x_2 \rangle \text{ that } [Gx_2])\]

is equivalent to, supposing a and b are the values of the variables,

\[\text{Tom bel of } \langle a \ b \rangle \text{ that } ^1[Fx_1 \& x_1 = x_2] \& \text{Tom des of } \langle b \rangle \text{ that } ^2[Gx_2] \& \text{TomSIV}(x_1 x_2, \text{that}^1, \text{that}^2)\]

and it makes no difference to Tom's psychological state whether or not a = b, or whether the values had been c and d instead. The psychologically relevant physical functional facts abstract from the particular identity of any res: the psychological theory is blind to the values of the variables, their identities and differences.

A narrow functionalist holds that all the psychologically relevant physical facts occur within the bodily envelope.
Since the concepts which Tom grasps are for the most part conceptions of the world around him, the narrow functionalist must hold that psychological theory is blind to the particular nature of Tom's environment, the properties of the things about Tom, as well as being blind to the particular denizens of that environment. For example, Putnam compares Tom on earth, who believes that his bath runneth over with water, and Twin Tom on Twin Earth, who believes that his bath runneth over with twater, a different stuff. Since Tom and Twin Tom are narrow replicas, Putnam postulates a common conception with a concealed indexical element - perhaps, the natural kind of which most of these are samples. Assigning the indexical one value on earth yields \( \text{H}_2\text{O} \) as the extension of "water". Assigning it a value on Twin Earth yields XYZ. The conception of water/determined by the narrow functional facts, and hence is blind to the particular properties of the environment which vary between earth, Twin Earth,...

The narrow functionalist must claim that all concepts a creature applies to its environment show this feature, even psychologically basic observation concepts like being red, being green, being sweet, being sour, etc. For example, compare SHRDLU - a robot equipped with a television 'eye' and a mechanical 'arm' who can recognise and manipulate blocks (cubes, pyramids, cylinders, etc.) employing such basic observation concepts as LINE, ANGLE (I use capitals to denote concepts) - and T-SHRDLU, a machine whose television eye and mechanical arm have been
replaced by leads to a computer running a block-world simulation program. Since SHRDLU and T-SHRDLU are narrow replicas, we suppose, the narrow functionalist must suppose there are common conceptions of LINE, ANGLE grasped by both SHRDLU and T-SHRDLU, and which apply both to the objects with which SHRDLU is perceptually and manipulatively interacting and to the inner states of the computer with which T-SHRDLU is interacting, though we would not say that those states were in any relevant sense linear or angular.

For functionalists wide and narrow suppose that extensions are determined by conceptual content plus the context in which the concept is exercised. In SHRDLU's case, this gives LINE, ANGLE, CUBE, PYRAMID, etc the extensions of lines, angles, cubes, pyramids, etc. because those are the features of SHRDLU's environment which are suitably placed in the input and output causal chains. And whatever constraints of counterfactual reliability etc. we place upon those causal chains in order that they be 'suitable', those constraints will be satisfied equally by the causal chains connecting T-SHRDLU to the relevant inner states of the computer with which it interacts. So those inner states are the extensions of T-SHRDLU's exercises of LINE, ANGLE and complexes of those inner states are the extensions of T-SHRDLU's exercises of CUBE, PYRAMID, etc.

However, a wide functionalist has more room for manoeuvre. His functional system extends beyond the bodily envelope. Suppose it extends to the gross physical bodies
the creature perceives and manipulates: cubes, pyramids, etc. in the case of SHRDLU. The wide functionalist has agreed that the functional system abstracts from the particular identities of the objects: the psychological life of SHRDLU is not affected if we substitute one pyramid for another, indiscernable under the concepts ANGLE, LINE. Does it abstract also from the particular properties "angle", "line" (I use " to denote properties) for which SHRDLU's sense is a reliable discriminative transducer? That is to say: Do SHRDLU and T-SHRDLU share the same psychological life, according to the wide functionalist? Do SHRDLU and T-SHRDLU grasp a common set of concepts? For if so, then the objects to which SHRDLU applies these concepts have the properties "angle", "line", whereas the states to which T-SHRDLU applies these same concepts do not possess those properties. So the common concepts are 'blind to' the difference between SHRDLU's world and T-SHRDLU'S world. If so, the wide physical functional system, like the narrow, abstracts from the particular first-order observable properties of the objects. But if not, if SHRDLU and T-SHRDLU do not share the same psychological life, if the physical functional system does not abstract from the observable properties of the entities with which the subject's interact, then SHRDLU grasps and exercises the concepts LINE, ANGLE, CUBE, PYRAMID, etc. which are conceptions of the properties "line", "angle", "cube", "pyramid" of the objects of SHRDLU's environment. And T-SHRDLU grasps and exercises the concepts
LINE', ANGLE', CUBE', PYRAMID', etc., which are conceptions of the properties 'line-', 'angle-', 'cube-', 'pyramid-', etc., properties of the inner states of a computer when running a block-world simulation program.

I shall argue, in section (iii), that the latter is the case, that SHRDLU and T-SHRDLU do not grasp and exercise the same concepts, that the wide functional psychological system includes the observable properties for which our senses are normally reliable discriminative transducers, and from these, therefore, the relevant physical functional system does not abstract. But first, in section (ii), I shall agree with Putnam that Tom and Twin Tom share the same concept (WATER), though I shall disagree over what that concept is, and over how it achieves a different extension on earth and Twin Earth.

Section (ii): Water and Twater

Putnam postulated a covert indexical element in the concept WATER: "water" means, roughly, "the predominant natural kind of these samples", or more likely, since there is no standard sample to which our usage defers, "the natural kind predominant in the colourless liquid more or less abundant in the environment of this neck of the celestial woods". Putnam combined this thesis with the view that Tom and Twin Tom speak different languages: English and Twin English. But, as Tyler Burge pointed out, postulating a covert indexical removes the reason for supposing Tom and Twin Tom speak different languages,
for it finds a common conceptual content for their utterances of "water" to express. However, it does follow that "water" does not have an extension, is not a predicate. Rather it is a concealed relation: "X is water" means, roughly, "x is of the same natural kind as y". "x is water" acquires an extension only if we are considering a particular utterance of it, in which a value is given to the concealed indexical, and are not considering it simply as a predicatable of English - the language common to parts of earth and parts of Twin Earth. No doubt Putnam's view that "x is water", as an English predicatable, has an extension led him to the view that earthlings and Twin Earthlings speak different languages.

I do not find Putnam's concealed indexical theory at all plausible. I accept from Putnam that the conceptual content of "water", on earth and Twin Earth alike, is something like: natural kind predominant in the colourless liquid more or less abundant in the environment. But I have omitted any indexical ("in this neck of the celestial woods") turning "x is water" into a concealed relation.

Twin Earth is a fantasy, and we are enquiring after our current concept of water prior to the discovery of a Twin Earth. The key point is that the requirement for an indexical varies from such fantasy to such fantasy. Suppose that when English speakers landed in the Americas they were settling in a land whose lakes and rivers ran with XYZ. Or suppose that lakes and rivers in the Southern Hemisphere are filled with XYZ. Or that when the earth's magnetic
field inverts, as I understand it does every so many thousand years, then $\text{H}_2\text{O}$ changes to XYZ, or vice versa as the case may be. Fantasies of course, but then so is Twin Earth. Now clearly, because these various fantasies draw the line dividing $\text{H}_2\text{O}$ from XYZ in different places they would motivate different indexicals concealed within "water", if the various Twin Toms' "water"-expressed beliefs are not to be chauvinistically declared false. Yet we are asked to believe that "water" really does contain a concealed indexical ready to deal with the Twin Earth fantasy. Why this favouritism; why doesn't it contain an indefinitely large plethora of concealed indexicals ready to deal with all the other fantasies too? (Don't say: Because we know the fantasies you have listed to be untrue, but Putnam's is still a'live' possibility. For there was a time at which we didn't know that my fantasies were untrue, yet we spake of water. And no doubt there are many such fantasies which we don't know to be untrue.) I suggest that there are no such indexicals concealed within "water".

There is no more to a concept than is manifest in our usage (wide function system). Nothing in that usage could reveal a concealed indexical poised to deal with one particular such fantasy, rather than alternative rival fantasies. What Putnam is really doing, I suggest, is spinning a plausible tale as to how our usage might extend, how our concept might develop, if his fantasy were true and we knew that it was true. Then we would
have a motive for inserting an indexical, an indexical which separates earth from Twin Earth (rather than the New from the Old World, the northern from the southern hemispheres, or ...).

Rejecting Putnam's theory of the concealed indexical, seems to land us back in the fire. If "water" on earth and Twin Earth expresses the same concept and that concept does not contain any concealed indexical, then, it seems, it will have the same extension on both earth and Twin Earth. What is that extension? It is not acceptable to say that "x is water" would have the null class as its extension because there are two natural kinds, $H_2O$ and XYZ, equally dominant in filling lakes and rivers. For that would make Twin Tom's "water"-beliefs massively false just because, unknown to him, earth is there too in the universe. And our "water"-beliefs would be subject to the hazard of there being a Twin Earth. Nor is it acceptable to take the conceptual content of "water" to be, roughly, natural kinds dominant in lakes and rivers. For, again, that makes Tom and Twin Tom's "water"-beliefs massively false, for each explicitly supposes water to be one natural kind.

A philosophical commonplace has it that concept expressed plus context of expression determine extension. Putnam's concealed indexical theory sought to show how concept plus context determine extension, by showing what feature of the context was the determinant - viz. the value of the concealed indexical. Having rejected that theory, we need some new theory to link the utterance to
the relevant features of its context. For, so far, we
don't know whether the context of my utterance of "water"
is this room now, Europe to the present, the earth past,
present and future, or the whole universe.

The missing factor, I suggest, is provided by the
communicative intentions of Tom and his fellows' language
community, and the communicative intentions of Twin Tom
and his language community. Their respective communicative
intentions do the work of Putnam's concealed indexicals.
And these communicative intentions are the exercise of
other concepts than WATER, concepts which employ explicit
indexicals. Tom intends to communicate with his fellow
earthlings about their common environment, he does not
intend to communicate with Twin Tom and his fellows about
their (Tom, Twin Tom, et al) common environment. And Tom
is not an eccentric member of his language community in
this regard. And of course, Twin Tom likewise. Now since
Tom and Twin Tom share the same mental life, wide as well
as narrow, they will share the same communicative intentions,
so far as conceptual content goes. But those communicative
intentions will include explicit indexical elements, de re
attitudes, which will have different values in their
different contexts - I intend to speak to you, and to
philosophical posterity. We can either say that indexicals
in communicative intentions determine the various contexts
which, together with the common concept WATER, determine
the different extensions. Or we can say extensions are
determined by concept plus context plus communicative
intentions.
If we did go on to discover that there has all along been a TwinEarth, we would not need to introduce indexicals into our concept WATER, the requisite indexicals already being in place elsewhere. But we would need a new concept TWATER, to talk about the contents of lakes and rivers over there, and this new concept would have a Putnamesque indexical element - not concealed but, in fact, the "t" of "twater". This new concept would not be the concept Twin Tom applies to what is in his lakes, rivers, etc. He, like us applies the concept WATER. Rather, Twin Tom's exercises of this new concept would apply our water, i.e. \( \text{H}_2\text{O} \), because of the different value of the indexical over there.

Given knowledge of Twin Earth and our now increased conceptual armoury, we would report TwinTom's beliefs thus:-

Twin Tom believes lakes are filled with twater... (1)

and not

Twin Tom believes lakes are filled with water ... (2)

And, of course, deleting the initial "Twin" from each, Twin Tom would report Tom's belief by (1) and not by (2). This is because we choose to preserve semantic content, when the different contexts and communicative intentions of believer and belief-reporter prevent us from preserving both conceptual and semantic contents.

We can confirm the relevance of communicative intentions by ringing changes on Putnam's example. Suppose earthlings and Twin Earthlings communicate by radio, no physical travel between planets occurs. Suppose earthlings
do not know the chemical composition of their water (in fact $\text{H}_2\text{O}$), nor Twin Earthlings that their water is XYZ. Imagine inter-planetary Gardener's Question Time, and a Twin Earth expert is advising an earth listener on the best regime for watering marrows. Ex hypothesi, the members of each population intend to communicate information about earthly and Twin Earthly matters to the aggregate of both populations. And there seems no reason to suppose that the concept each population associates with "water" has changed from Putnam's original example, where each population was ignorant of the existence of the other. But the truth conditions of their utterances are now different, because they intend their utterances to be assessed for truth value on earth and Twin Earth indifferently. So the relevant context has been changed by the change of communicative intentions. And the extension of the term "water", as used on both earth and Twin Earth, is now $\text{H}_2\text{O}$ and XYZ. Both populations are now wrong in thinking that water is one natural kind, and likewise, presumably, in thinking that marrows are one natural kind - suppose if you did indeed pour a bucket of XYZ over an earth marrow it would shrivel up.

Now suppose we add to the above that earthlings generally (or perhaps earth experts to whom earth usage defers) believe that water is $\text{H}_2\text{O}$, Twinearthlings that water is XYZ, but neither party has yet twigged the real difference between earth and Twin Earth. We have a trilemma. Either the extension of "water" is $\text{H}_2\text{O}$ and not
XYZ. In which case what an earth expert says a Twinearthling should do with his marrows is false, $\text{H}_2\text{O}$ being fatal to marrows on TwinEarth, even though the expert intended to help the listener, and even though his utterance furthered rather than hindered that intention. Or the extension of "water" is XYZ, and similar remarks apply concerning Twin Earth advice to earthlings. The cost of these first two lines is that interplanetary communications are massively false and yet massively successful. Or, thirdly, the extension of "water" is both $\text{H}_2\text{O}$ and XYZ. But now earth-chemist-to-earth-chemist communications are massively false, because false of water on Twin Earth, as are Twinearth-chemist-to-Twinearth-chemist communications, being false of water on earth. There is no need to try to resolve this trilemma: their intraplanetary communicative intentions clash with their interplanetary communicative intentions.

Finally, let me note that my theory makes the following both true!

Twin Tom believes that lakes are filled with twater. Twin Tom applies the concept WATER to lakes. Yet we need not conclude that WATER = TWATER. For, given the communicative intentions of Twin Tom and his peers, the concept WATER applies to twater, and the concept TWATER to water!
Section (iii): Wide functionalism and observation concepts

Wide and narrow functionalists alike feel relatively comfortable saying that Tom and Twin Tom share the same mental life, exercise the same repertoire of concepts, because Twin Earth looks, sounds, feels, tastes and smells from Twin Tom's point of view exactly like earth does from Tom's point of view. But SHRDLU (the robot manipulating blocks) and T-SHRDLU (the computer manipulating another computer's block-simulation program) are more unsettling. Water is, of course, observable stuff, but we lack a sense which is a reliable discriminative detector of water – pace the claims of cranks with hazel twigs – like our sight is a reliable discriminative detector of red – reliable relative to certain readily satisfied boundary conditions, which are usually referred to as the 'normal conditions' for seeing. RED is an observation concept for normally sighted humans.

SHRDLU's observation concepts are LINE and ANGLE. But T-SHRDLU does not 'observe' a world of blocks, and its sense is not a reliable discriminator for lines and angles. Hence we feel more uneasy, than in the case of Tom and Twin Tom, in attributing the same conceptual repertoire, and the same exercises of the same concepts to T-SHRDLU as to SHRDLU. The narrow functionalist has no choice. As Fodor remarks:—

if there is a narrow notion of [conceptual] content, it must be determined independently of [semantic] interpretation: it must be determined by functional role

So SHRDLU and T-SHRDLU, realising the same narrow functional
states, grasp and exercise the same concepts. The wide functionalist may be tempted to follow this route too. Although the wide functional systems which realise SHRDLU and T-SHRDLU are constituted by different first-order properties (lines and angles in the one case, states of a computer in the other), the physical and psychological functional systems abstract from this first-order difference, as they abstract from the identities and differences of particular res.

However, I do not recommend this route to the wide functionalist, and regard it as an advantage of wide functionalism over narrow that there is an alternative: We can include the observational properties in the relevant physical facts, as part of the physical functional system. These physical facts determine SHRDLU's grasp of the concepts LINE, ANGLE and of the derived concepts CUBE, PYRAMID, etc. And the different physical facts about T-SHRDLU determine a different conceptual repertoire, concepts apposite to the internal states of a computer running a block-simulation program.

A reason why I do not recommend wide functionalists to follow the path which is forced on narrow functionalists, is that the denizens of the extra-cranial world would not then instantiate concepts in a sufficiently absolute, mind independent way. It offends against my naturalistic prejudice. We say that the environment of SRDLU consists of lines, angles, cubes, pyramids, etc., and that inside the computer T-SRDLU plays with there are no such things.
And contrariwise, if we were able to identify the computer states which T-SHRDLU causes and responds to, we would find such concepts do not apply in the locale of SHRDLU. But, according to narrow functionalism and this current view of wide functionalism, SHRDLU and T-SHRDLU exercise the same repertoire of concepts, and the beliefs of each about its respective environment are true. So from the points of view of our SHRDLUs these two environments share the same set of properties, instantiate the same set of concepts.

Perhaps SHRDLU's interactions with blocks etc., and its computational processes, are not rich enough to furnish it with the concepts BLOCK, etc.? But if not, suppose its interaction with its world and its computational processes are so enlarged that there is no doubt that it grasps the requisite concepts if we do. It still seems that T-SHRDLU and its block-simulation program could be similarly enlarged.

What goes for SHRDLU goes for me too. However full a report is given of the relevant physical facts about me, if that report abstracts from, as not being psychologically relevant, the identities of particular res and the particular observable properties of my environment, then there could be another creature, inhabiting a radically different neck of the universe, to whom the same physical theory applies, and hence, if that theory contains all the psychologically relevant physical facts, that other creature and I share the same mental history. Thus, although I, from my point of view and with my communicative
intentions, may describe his neck of the woods as not
having the properties and not instantiating the concepts
which mine does, he from his point of view and with his
communicative intentions applies those very concepts and
recognises those very properties in his neck of the woods,
and denies them of mine. And both of us are correct.
There is no God's-eye privilege which I can warrantedly
claim but he cannot, and who dare advance the thesis that
really, but unknowably to either of us, one of us is
correct and the other wrong. Thus there is no mind-
independent absolute sense in which one or other of the
environments, but not both, really does instantiate one
or the other set of concepts, really do have one or the
other set of properties. From my point of view my neck of
the woods instantiates a certain set of concepts and his
does not, but instantiates some other set of concepts.
But from his point of view his neck of the woods instantiates
the first set of concepts and my neck of the woods some
other set of concepts. Truth (whether a given entity
instantiates a given concept) is relative to a cognitive
point of view.

Is this really only our old WATER/TWATER relativity
of section (ii)? Then we said that, given their respective
communicative intentions, Tom and Twin Tom both correctly
apply the same concept WATER to the stuffs in their
respective lakes and rivers, and yet the following are
both true belief-reports:–

Tom believes rivers run with water
Twin Tom believes rivers run with twater
So, it might seem, given their respective intentions to talk about their respective environments, both SHRDLU and T-SHRDLU apply the same concepts LINE and ANGLE to their respective environments, and yet the following are both true belief-reports

SHRDLU believes an entity presents three lines forming three angles.

T-SHRDLU believes an entity presents three lines' forming three angles'.

where 'lines'' and 'angles'' are properties of the internal states of its attendant computer.

What are the graspable concepts of LINE and ANGLE common to both SHRDLU and T-SHRDLU? Since we take these to be conceptions of properties for which the creature's senses are reliable discriminators, let us follow Peacocke and suppose that, for each such observation concept, there is a sensational property. These are the properties 'line' and 'angle' of SHRDLU's and T-SHRDLU's televisual fields. We are now supposing that SHRDLU's and T-SHRDLU's environments are, to them, lookalikes.

(I use 'line' etc. to denote a property of their televisual field, because I am already using 'line' etc. to denote the corresponding property of SHRDLU's environment, and 'line' to denote the corresponding property of T-SHRDLU's environment.) Now we can define the common conceptions LINE etc. as follows:-

\[ x \text{ instantiates } \text{LINE} =_{df} x \text{ is presented by a line'' in the televisual field, in normal perceptual circumstances.} \]

\[ x \text{ instantiates } \text{ANGLE} =_{df} x \text{ is presented by an angle'' in the televisual field, in normal perceptual circumstances.} \]
Following Peacocke, we must distinguish the **definitional priority** of line" etc., to LINE etc., from an **epistemic priority**: we do *not* suppose that SHRDLU or T-SHRDLU grasp the concepts LINE" etc. Now we can report the **contingent facts** about SHRDLU

\[(x)(x \text{ is a line} \leftrightarrow x \text{ is presented by a line}" \text{ in the televisual field, in normal perceptual circumstances})\]

Etc.

And the **contingent facts** about T-SHRDLU

\[(x)(x \text{ is a line}' \leftrightarrow x \text{ is presented by a line}" \text{ in the televisual field, in normal perceptual circumstances})\]

Etc.

If the above is accepted, then SHRDLU/T-SHRDLU do *not* point to any worrying mind-independent **ontological relativity**. Supposing their respective programs and capacities expanded until their domains overlap, SHRDLU will apply a given **concept** to a given entity, which **concept** T-SHRDLU will deny of that entity, and vice versa. But this does *not* now imply that the **properties** of that object are mind dependent - for it can be a line without being a line', or vice versa. So, if the above is accepted, the wide functionalist can abstract from the particular observation properties: line and angle, or line' and angle'. He is *not* thereby retreating to narrow functionalism. He need not concede that the narrow proper part of the wide functional system is sufficient to realise the conceptual contents LINE and ANGLE, and their computational, semantic-value-preserving, employment.
The story of this section so far. We are exploring the option open to a wide functionalist of abstracting from the particular observational properties encountered by SHRDLU and from those encountered by T-SHRDLU. I.e., the option, in the physical theory which determines the psychological facts, of quantifying over the observation properties, and, in the psychological theory, using the sensational properties line', angle'' as primitives, and defining the concepts LIKE, ANGLE, which are then only contingently related to whatever are, in fact, the observational properties for that creature.

However, I do not recommend the option thus developed to the wide functionalist. For it raises the spectre of a kind of inverted spectrum, and the mere possibility of such spectral inversion is anathema to my naturalism.

We have defined objectural concepts in terms of sensational properties thus:-

\[ x \text{ instantiates } \text{RED} =_{df} x \text{ is presented by a red'' area of the televiual field, in normal perceptual circs.} \]

\[ x \text{ instantiates } \text{GREEN} =_{df} x \text{ is presented by a green'' area of the televiual field in normal perceptual circs.} \]

Compare SHRDLU for whom

\[ (x)(x \text{ instantiates } \text{RED} \leftrightarrow \text{Red } x) \]

\[ (x)x \text{ instantiates } \text{GREEN} \leftrightarrow \text{Green } x) \]

in normal perceptual circumstances, of course, with I-SHRDLU whose televiual eye has had a 'colour-reversing' lens inserted: the lens has the effect of inverting the apparent colours of things rather as the negative of a colour photograph compares with a positive print. For I-SHRDLU we have:-
(x)(x instantiates $\text{RED} \leftrightarrow \text{Green} \ x$)
(x)(x instantiates $\text{GREEN} \leftrightarrow \text{Red} \ x$)

Assuming that the psychologically relevant physical facts quantify over the observational properties red' and green', then the psychologically relevant physical facts about SHRDLU are the same as those about I-SHRDLU. True, at the level of first-order facts, I-SHRDLU has an extra part, but that part does not register in the psychologically relevant physical facts: those facts merely relate some pair of observational properties to a pair of internal states. Since SHRDLU and I-SHRDLU share the same psychologically relevant physical facts they also share the same psychological histories. I.e., when SHRDLU is enjoying an $x$ presented redly'', so is I-SHRDLU! And green'' presentations likewise. Inserting the lens has made no difference to I-SHRDLU's psychological history compared to that of SHRDLU. There is no spectral inversion.

However, there arises a possibility of metaphysical spectral inversion. Given a red cube and a green pyramid viewed by both SHRDLU and I-SHRDLU, which sensational property, red'' or green'', does the cube cause in them, and which the pyramid? One wants to say that SHRDLU sees the cube redly'' and the pyramid greenly'', whilst I-SHRDLU sees the colours reversed. But they are psychological replicas, so we must say the same for each. And clearly the choice is arbitrary.

The root of the problem is that the physical functional theory has been pitched at a level of abstraction
which results in that theory being isomorphic with itself under a mapping which is not identity. Such a possibility clashes with my brand of naturalism: naturalism being the view that physical facts determine (non-reductively) psychological facts. But our physical functional system provides no explanation of why the one sub-system realises the sensational property red" and the other realises the sensational property green"; rather than vice versa. That the one physical sub-system relises red" and the other green"; rather than vice versa, is a metaphysically brute psycho-physical correlation. God had a free choice, and could have inverted the correlation, it seems.

There are some brute psycho-physical facts. As Nagel points out, I could have complete mastery of the wiring diagram for a bat, but still not know what it is like to be a bat. It is a metaphysically brute fact that being wired up like this (the build-a-bat blueprint) realises experiences like that (what it is like to be a bat, whatever that may be). A naturalist has to claim, however, that being wired-up like this just is metaphysically sufficient to having that subjective experience. The present problem is that being wired-up like this is not sufficient to having those visual experiences. The same physical functional system could, had God so chosen, have realised the complementary sensational experiences, without change of psychological function.

However, the problem vanishes if we reverse the conceptual priority; if we take RED and GREEN as the
primitive concepts, and RED'' and GREEN'' as the defined concepts thus:-

$$\text{Red''} x = _{df} x \text{ is how a red object normally presents in perception}$$

$$\text{Green''} x = _{df} x \text{ is how a green object normally presents in perception}$$

Now the physical functional system which realises red'' sensations extends out to televisual contact with red objects, and the physical functional system which realises green'' sensations to green objects. And most importantly, the physical functionally related facts which determine the psychological facts do not abstract from, but rather include, the observational properties red' and green'. So God is not called upon to choose one or another psycho-physical correlation.

Now a wide functionalist allow that I-SHRDLU has an inverted spectrum relative to SHRDLU; that green objects are presented by red'' areas of its televisual field, and red objects by green'' areas? It depends. Yes, if I-SHRDLU was a normally functioning SHRDLU into whose eye the colour-inverting lens was inserted. For the insertion of the lens changes the functional relations, in that red objects are now classified as being of the same colour as green objects are remembered as having been. And green objects are classified as being of the same colour as red objects encountered earlier. Whereas SHRDLU classifies red objects as being the same colour as red objects encountered earlier, and current and earlier green objects likewise. No, if I-SHRDLU was 'born' in this condition -
i.e., if not only are the first-order physical internal states which constitute seeing redly" and seeing greenly" inverted as between SHRDLU and I-SHRDLU (these states are identified in accord with the definitions above as those normally caused by red and green objects respectively), but also the first-order physical states which constitute memories of red objects and memories of green objects are inverted as between SHRDLU and I-SHRDLU. The two first-order inversions 'cancel out' in that they fail to produce a psychologically relevant functional difference.

All naturalists hold that the psychological is somehow supervenient upon the physical. Narrow functionalists must take sensational properties as primitive, and objectural properties as defined, in the psychological theory. But now inversion of the subjective spectrum relative to those objects which are in fact red and green is a possibility, since the causal chains from external objects to narrow states may be interchanged external to the narrow system. And supervenience plus narrow functionalism requires that Tom and Twin Tom, enjoying the same narrow states of the same narrow system, enjoy the same sensations. But such narrow functionalism plus psycho-physical supervenience leaves it as a brute, magical psycho-physical correlation that those narrow physical states realise those sensational properties, rather than realise their complementary sensational properties.

On the other hand, if the psychological is supervenient upon a wide physical functional system, and if that physical
system is not blind to but includes things being red and things being green, then there is no such brute magical psychophysical correlation about supervenience. Those physical states which involve red objects realise red' sensations, and those physical states which involve green objects realise green'' sensations. No matter that inserting a lens whilst leaving internal wiring intact (memory connections), will connect red objects to green'' sensations and green objects to red'' sensations. For red objects are still connected to the old memory states, and green objects likewise. If the memory to perception connections are also reversed, then there is no inversion of sensations relative to objects.

Does this not accord with our intuitions? I can imagine waking up to find the colours of things have apparently inverted relative to what I remember them to have been. But, post Wittgenstein, to imagine waking up with colours apparently inverted and memories of colours also inverted, is not to imagine any change at all.

My wide functionalism thus includes the semantic contents of certain internal representational states. The functional theory assigns to certain representational states the semantic content red', and to certain others the semantic content green'. Fodor is persuasively pessimistic about the prospects for a natural psychology, a psychology which seeks to give the semantic contents of mental representations in a scientific, projectible vocabulary. (A psychology which characterises the semantic
content of "salt" as NaCl, rather than as what Granny
likes with herring.) Briefly, he points out that it was
chemists, not psychologists, who discovered that the
extension of "salt" is sodium chloride, and, he asks,
what is the science dealing with "undertakers", "umbrellas",
... ?

Points taken, but it may be that for wide functionalism
we need only a natural psychology of observable properties,
properties for which our senses are reliable discriminative
transducers, in normal circumstances, together with a basic
conception of an enduring physical object as the bearer
of such properties. Perhaps these properties will be
revealed by the physical sciences to be secondary properties.
If so, they must be rejected. I discuss the primary-secondary
quality issue further in Appendix B.

I do not know how to draw a sharp line between those
properties for which our senses are reliable discriminators,
in normal circumstances, and those properties for which
they are not. I.e., between those properties which are
eligible for inclusion in the psychologically relevant
physical facts, and those which are not. Observable
properties include being red and being square. Should we
add being water to the list? I don't want to, but water
is certainly observable, and although other liquids are
colourless etc., the normal circumstances in which we have
evolved, and the normal circumstances in which we learn
language, do not include lakes, rivers, puddles, rainfall,
taps-supplying etc. these other liquids. Perhaps water
is disqualified because our best possible individual or collective sense-based judgement (no chemical experiments) that something is water could be mistaken, so it is not an observable property like being red or being square? But no doubt in suitably bizarre and contrived circumstances our best possible individual or collective sense-based judgement that something is red or square would be mistaken, and that the prevailing circumstances were bizarre in this way need not be itself observable. Perhaps the answer is that we should await the development of wide functional theory to show us what properties of our environment are its observable properties, and not to seek a criterion dividing observable from other properties prior to that theory.
Lewis is a narrow functionalist who, I think, gets into trouble over the possibility of an inverted spectrum. Here is a statement of Lewis's narrow functionalism, cast in the form of a comment on an imaginary madman, Heimson, who believes himself to be the real Hume.  

Heimson may have got his head into perfect match with Hume's in every way that is at all relevant to what he believes. If nevertheless Heimson and Hume do not believe alike, then beliefs ain't in the head! They depend partly on something else, so that if your head is in a certain state and you're Hume you believe one thing, but if your head is in that same state and you're Heimson you believe something else. Not good.

The main purpose of assigning objects to attitudes is, I take it, to characterise the states of the head: to specify their causal roles with respect to behaviour, stimuli, and one another. If the assignment of objects depends partly on something besides the state of the head, it will not serve this purpose. The states it characterises will not be the occupants of the causal roles.

A narrow functionalist cannot avoid the possibility of an inverted spectrum. Heimson and Hume might be perfectly matched head cases, but that internal state which in Hume is causally triggered by fire engines, pillar boxes, etc. is, in Heimson, causally triggered by green-painted fire engines, green-painted pillar boxes, etc. And that internal state which in Hume is causally triggered by grass, conifer for ests, etc. is, in Heimson, causally triggered by red-painted grass, red-painted conifer for ests, etc.

Of course, Heimson is special because Lewis has postulated that he shares each token belief of Hume's. Usually, it is sufficient for inversion that whatever internal state is normally caused by sight of red objects is, in the invert, caused by sight of green objects, and
vice versa, whatever internal state is normally caused by sight of green objects is, in the invert, caused by sight of red objects. There is no requirement to paint the objects which the invert sees, for the inversion lies in the causal routes from properties of objects to internal states - say, a spectrum inverting dust in the inverts atmosphere, say a special coating on his eye immediately on the 'world' side of whatever boundary the narrow functionalist draws.

Lewis agrees that subjective spectral inversion is possible and writes:

I would say that there is a good sense in which the alleged victim of inverted spectra sees red when he looks at grass: he is in a state that occupies the role of seeing red for mankind in general. And there is an equally good sense in which he sees green: he is in a state which occupies the role of seeing green for him, and for a small subpopulation of which he is an unexceptional member and which has some claim to be regarded as a natural kind. You are right to say either, though not in the same breath. Need more be said?

This much more: What, then, is it like to be him? My criterion for psychological realism (Chapter Two) was that the physical facts realise only one psychological history - though that history may be intermittent and disjointed (Chapter Four, Appendix B). A subject cannot literally be in two minds at once. But Lewis's victim of an inverted spectrum is, it seems, literally in two minds at once. For he is a member of both populations at once.

So much the worse for narrow functionalism, say I. A wider functionalism agrees with Lewis that the purpose of assigning contents to attitudes is to characterise
states centered on the head: to specify their causal roles with respect to behaviour, stimuli, and one another. But wide functional states which embrace objects and their observable properties can hope for contents which are unique enough both to satisfy our intuitions that there is something it is like to be that creature, and to provide the intersubjectively graspable meanings of our utterances.
I have argued for a wide functionalism which is not blind to observable properties. To do so is to take those properties as primary qualities, as pukka properties of the objects perceived. For if we treat them as secondary qualities, on the model of, say

\[ x \text{ is funny} \equiv \text{df } x \text{ is disposed to cause amusement in humans,} \]

then we get

\[ x \text{ is red} \equiv \text{df } x \text{ causes red'' sensations in humans in normal perceptual circumstances} \]

\[ x \text{ is green} \equiv \text{df } x \text{ causes green'' sensations in humans in normal perceptual circumstances,} \]

and the resulting functionalism is wide, but blind to observable properties. For Tom and Twin Tom can both realise the same wide functional system, so characterised, and the properties of objects which respectively cause red'' and green'' sensations in Tom differ as radically as we please from those which do the same for Twin Tom.

Whether a given observable property turns out to be a primary or a secondary quality is, I think, an a posteriori question. It may well be that, from a physicist's point of view, the property red' which gives objects the power to cause red'' sensations in humans in normal circumstances is an uninteresting disjunctive property. For example, we could build a machine which bleeps either when a microphone is activated at a frequency of 440c.p.s. or when a photocell is illuminated by 10 flashes per sec.
Calling such a machine an X-detector, it turns out that the property of having X is the disjunctive property of either producing vibrations at 440c.p.s. or flashing at 10 beats per sec. Not a property of any interest to physical theory. Now it might turn out that our senses are reliable discriminators, in normal circumstances, for such uninteresting disjunctive properties - indeed, given the chemical complexity of foodstuffs and the comparative paucity of my senses of taste and smell, I'd put a modest wager on it.

Furthermore, it may be that only in normal visual circumstances do objects having $PvOvR$ cause us to see red. If so, and if 'red' is a primary quality, then we may discover that, in unusual circumstances, red objects do not look red. Our pretheoretical colour judgements stand to be corrected by science, if 'red' indeed turns out to be a primary quality. Of course, the more the havoc wrought upon our pretheoretical colour judgements, the greater the pressure to treat colours as secondary qualities.

So I had better not give hostages to empirical fortune by claiming that wide functional theory incorporates all observable properties, that none will turn out to be secondary qualities. But I'd be surprised if we have not evolved to reliably detect observable primary qualities on the whole.

My view is that it is an a posteriori matter whether a given observable property is a primary or a secondary quality. McGinn disagrees. He writes¹¹:-
The key point is that it is an a priori matter whether some quality is primary or secondary, i.e., whether it is to be analysed in terms of a disposition to produce experiences: we can tell, by reflecting on the concepts, that (e.g.) red is secondary and square is primary.

McGinn offers three kinds of argument for his claim. The first, the main theme of his book, is a transcendental argument that perception of an observer-independent world of objects bearing primary qualities must also involve secondary qualities. I do not find this argument convincing, but I shall not comment further. His second argument is a claim that our knowledge of secondary qualities is incorrigible. He writes:\footnote{12}

> the identification of the secondary qualities of a thing is incorrigible. The reason is that we know incorrigibly the kind of experience an object of perception produces, and on the dispositional analysis this is to know which secondary qualities the producing object has... this needs some qualification to make room for the distinction we customarily draw between real and apparent secondary qualities; but this does not seriously disturb the incorrigibility claim because it can either be made to apply at the level of the community or we can consider the simplified case of the solitary perceiver...

Thirdly, we have a priori knowledge of quasi-logical relations between secondary qualities, typified by our a priori knowledge that nothing is red and green all over at the same time:\footnote{13}

> each sense brings with it a system of secondary qualities exhibiting quasi-logical relations which arise from the subjective nature of the experience involved, these relations not being reducible to anything objective.

Of course, we also know a priori that nothing can be round and square at the same time. But that was hard won a priori knowledge resulting from difficult analysis of the
objective properties of being round and of being square. By contrast, our a priori knowledge that nothing is red and green is easy, since it follows directly from our immediate knowledge that we cannot imagine an experience as of something both red and green, and that is enough to establish the truth about objects not being red and green. On the other hand, if red' and green' were non-relational properties of objects, we could not know a priori whether or not something could be both red and green.

I think these arguments of McGinn fail, and fail for a philosophically important reason. There is no more to our concepts RED, GREEN, etc. than is manifest in our usage. And McGinn's intuitions may reflect his conception of redness, but usage gives no warrant to choose between a conception of redness as a dispositional secondary quality, and a conception of redness as a primary, manifest quality. Philosophers tend to precisify such concepts, and hence their intuitions and philosophical theories reflect that precisification. But if one set of philosophers precisify in one direction (dispositional property) and another precisify in another direction (manifest property) and both claim to be speaking for language users as a whole, then the issue cannot be settled by appeals to the a priori knowledge of the ordinary language user. For the ordinary language user does not distinguish a priori from a posteriori knowledge with anything like the required precision, nor does he distinguish epistemic from ontological modality. So when an ordinary language user
agrees that nothing can be red and green all over, and that the community's judgements cannot be wrong, he is interpreted by one set of philosophers as voicing a priori knowledge of the ontological necessities of his own sensory constitution, and he is interpreted by the other set of philosophers as voicing epistemic truisms drawn from a background theory so well entrenched that he does not regard it as threatened by experimental enquiry.

Cases which would force the generality of language users to further determine their colour conceptions, in one direction or the other, have not arisen in practice. This is why, I claim, it is an a posteriori matter whether red etc. are primary or are secondary qualities. And, I conjecture, common usage takes properties to be primary until the cost of doing so becomes exorbitant - because too many pretheoretical colour judgements are reversed. So when I claim that it is an a posteriori matter whether a given observable property is a primary or a secondary quality, I do not mean that there is now a determinate fact of the matter, which fact remains to be discovered a posteriori. More likely, it is now indeterminate, but with bias towards primary, and a posteriori which way the conception would develop were cases to arise in practice which required us to make up our collective mind.

To see how we could be forced off the fence, suppose Martians land and integrate with us. Suppose their visual sense organs, instead of having just three kinds of colour receptor like ours, have some 30,000 differentially
sensitive kinds of colour receptor (as our ears have some 30,000 differentially sensitive frequency receptors - no doubt this physiological difference between our sight and our hearing accounts for our inability to see a surface as having more than one colour and our ability to hear, after training, each violin, the viola and the cello issuing from the same gramophone speaker). It should come as no surprise that if we illuminate a white surface with equal brightnesses of red and green lights (the lighting being concealed from the Martian) the Martian will say, like us, that the surface is a yellowish white, but, if he concentrates, he can also see it as uniformly red and green all over at the same time, that he can switch from the one colour-gestalt to the other.

Suppose that Martians had no problem learning our colour language and in joining in our colour ascriptions. They simply need to take a quick and careless look, not to study the colour too closely, to make the same colour judgements as humans. And suppose also that the occasions upon which they can make their distinctive multi-colour judgements are rare: most natural terrestrial objects produce such a cacophony of wavelengths that the Martian cannot resolve them into some few components, white screens are like acoustic 'white noise', and they can do their party trick only in contrived circumstances like white screens illuminated by a few monochromatic beams. But suppose some of these cases are important to us human beings, so we cannot simply sweep them aside as curiosities. Perhaps nuclear fuel rods glow yellowish white in use, but
this yellowish white tends towards a composition of red and green monochromatic components as they become critical.

One way in which our language could adapt to the new situation, McGinn's way, would be to say that the rods were yellowish white to humans, and red to Martians and also green to Martians, and to retain the thesis that nothing can be both red to humans and green to humans. Thus we now have two systems of secondary qualities, red etc. to humans, and red etc. to Martians. Colours are relations, and objects are not red etc. per se. Thus the property red-for-humans' is the property of causing red'' sensations in humans. And the property red-for-Martians' similarly. The resulting wide functional theories of humans and of Martians will be blind to the manifest properties of the objects of perception, at least so far as red' is concerned.

Despite McGinn's confidence, we might jump the other way. We might say that Martians have more accurate colour vision than we have, as we have better colour vision than cats. In most cases they agree with us, and those cases in which they differ from us matter to us in that the Martians gain useful information about the world which we lack. Our colour vision, we decide, is defective and hence we were wrong, as a community, in claiming that certain yellowish white objects were also not red and not green. We already suppose that dogs can hear sounds that we cannot, and cannot even imagine (I cannot imagine a sound of even higher frequency than the highest note that
I can hear). So, that I cannot imagine something being red and green does not force my logical hand.

However, the above is merely a thought experiment. It is an experiment which would invite our general usage to precisify either along McGinn's line (a priori secondary quality) or not (primary quality). But since it takes such a hypothetical thought experiment to force the issue, we cannot claim that actual communal usage manifests one of the other of the rival conceptions, whatever the individual intuitions of individual philosophers.

It is therefore an a posteriori matter whether or not there will come to be a fact of the matter as to whether the observable property of red' is a primary or is a secondary quality, and an a posteriori matter what the fact of the matter comes to be, if it does come to be.

Finally, let me note that some philosophers have claimed that red' is a secondary quality because it is sense-specific. Only those who can distinguish red objects by sight have the concept RED. A man born blind, though he may somehow successfully sort things by colour, say he can do so by touch, cannot have our colour concepts. Hence they are secondary qualities, for by contrast, being square is not so tied to one sense modality, nor to any sense modalities in particular. An intelligent bat could have our shape concepts thanks to its sonar sense. I reject this supposedly acid test, for the reasons given above. De facto all colour discrimination is done by sight. But whether this is so de jure requires the invocation of a counterfactual thought experiment, and such are powerless to determine the issue.
Footnotes to Chapter Six

1  Putnam "The Meaning of 'Meaning!'", in Putnam (1)
2  Tyler Burge "Other Bodies" in Woodfield (1)
3  Common to Putnam and Fodor (in "Danish disContent" in Butterfield (1)
4  Fodor (2) p106
5  The functional system described by Loar (1), section 4.4 p67ff is such a system. The input conditions have the form

   If p is true, 0(p,z,t) and..., then z believes at t that p

   where p is a proposition that an entity has a certain observable property, and 0(p,z,t) means that z is observationally related to p at t. So the actual input conditions of Loar's functional theory will employ particular observational concepts: e.g.

   If it is true that this is red, 0(that this is red,z,t) and..., then z believes at t that this is red.

   If such a theory were formulated for SHRDLU's psychological states, it would not apply to T-SHRDLU.

6  Peacocke (2) Chs 1 and 2
7  Compare ibid. p42/3
8  Fodor (2)
9  Lewis (2) p142-3
10 Ibid. p128
11 McGinn (3) p144
12 Ibid. p47
13 Ibid. p35
14 Who are these other philosophers? One is Kripke (2) ft. 71 where he argues that yellowness is a manifest and not a dispositional property.
CONCLUSION

Let me draw the threads of this discussion together to reach a conclusion.

(A) In the case of our self-ascribed folk-psychology we have a warrant to apply the conceptual contents of sentential indices to the states so indexed, provided the sentences of our home language do have determinate conceptual contents (Chapter Two, sections (ii) and (v)). And we have reason to conclude that those sentences do have determinate conceptual contents (Chapter Three).

(B) Narrow functionalism cannot account for these conceptual contents because:

(i) Any narrow state indexed by "Boise is a city" could, it seems, also be indexed by "'Boise' is stored at the address CITY", a sentence with a different conceptual content, whereby neither content can be ascribed to the state so indexed (Chapter Two, section (iii)).

(ii) Field's narrow functionalist semantics for classical logic does not assign to "\( \exists \)" the sense "there is" rather than the sense "there may be", with variable modality (Chapter Four, section (iv)).

(iii) A narrow functionalist cannot show that the quantifier rules and inductive rules preserve designated semantic values, which last is a necessary condition of the indexed states being the psychologically relevant functional states (Chapter Four, section (iv)).

(iv) A narrow functionalist is left with metaphysically brute psycho-physical correlations between
sensational properties and physical inner states, with no possible explanations as to why the correlations are not permuted (Chapter Six, section (iii)).

(C) Wider functionalism, on the other hand, does not encounter or can solve these problems (references as above, plus Chapter Two, section (iv) and Chapter Four, sections (ii) and (iii)).

(D) However, despite the above reasons in favour of a wide functionalism, it still seems that we could take any episode of Tom's psychological history and give it a narrow functional explanation, an explanation which does not involve attributing any res to Tom's attitudes, and an explanation which can be interpreted realistically because Tom would avow the various content clauses. For example, suppose a 'wide' explanatory complex of the following form applies to this episode of Tom's psychological history:

\[(3x_1x_2x_3x_4)(\text{Tom per of } \langle x_1 \rangle \text{ that } [Fx_1] \land \text{ Tom bel of } \langle x_1x_2x_3x_4 \rangle \text{ that } [Gx_1x_2x_3x_4] \land \text{ Tom des of } \langle x_3 \rangle \text{ that } [Hx_3] \land \text{ Tom int of } \langle x_4 \rangle \text{ that } [Ix_4]) \]  

\[
\text{Now supposing } a, b, c, \text{ and } d \text{ are in fact the relevant res, Tom's behaviour is equally explained by:--}
\]

\[
\text{Tom per of } \langle a_1 \rangle \text{ that } [Fx_1] \land \text{ Tom bel of } \langle a_1b_2c_3d_4 \rangle \text{ that } [Gx_1x_2x_3x_4] \land \text{ Tom des of } \langle c_3 \rangle \text{ that } [Hx_3] \land \text{ Tom int of } \langle d_4 \rangle \text{ that } [Ix_4] \land \text{ Tom SIV(}x_1-4, \text{ that } 1-4) \]  

But, it seems, it makes no difference to our ability to explain Tom's behaviour if we treat this whole episode as if Tom were hallucinating a, b, c and d. Then we would have the explanation:-
Tom per that \(1[F_{y_1}] \) & Tom bel that \(2[G_{y_1y_2y_3y_4}] \)
& Tom des that \(3[H_{y_3}] \) & Tom int that \(4[I_{y_4}] \)
& Tom SIV\(^1\)(\(y_{1-4}\), that\(^{1-4}\)) \(\ldots\ldots\) \(\ldots\ldots\ldots\ldots\ldots\) \(3\)

I have changed the indexicals from \(x\)'s in (2) to \(y\)'s in (3) because the indexicals in the content clauses of (2) are required, subject to certain conditions, to retain the semantic contents of the indexicals in the sentences to which Tom is disposed to avow, whilst the indexicals in the content clauses of (3) are required to have the same linguistic meanings as Tom's indexicals, so we should not expect the indexicals in the content clauses of (3) to be the same as their counterparts in the content clauses of (2). (Indeed, the indexicals in the content clauses of (3) may even acquire values in the context of utterance of (3), but those values are irrelevant to the psychological explanation of Tom's behaviour.)

Now it seems that, given any explanatory complex involving quantification and the quasi-binding of variables in the manner of (1), we could construct a 'narrow' explanation in the manner of (3), which narrow explanation is equally good at explaining the subject's behaviour.

(E)Conclusion: Although any single episode can be explained narrowly, and genuine hallucination can only be explained narrowly, assigning conceptual contents and rules of inference to such narrow attitudes presupposes that sufficient episodes are explained in the manner of (1).

The appeal of narrow functionalism is the realisation that any episode may be treated narrowly without explanatory loss. The fault of narrow functionalism is the view that all episodes can be so treated together.
BIBLIOGRAPHY

Of works mentioned

Benacerraf P (1): "What Numbers could not be": Phil. Rev. 74 (1965)


Burge T (1): "Reference and Proper Names": J. Phil. 70 (1973)

Burge T (2): "Beliefs de re": J. Phil 74 (1977)


Davidson and Hintikka (eds) (1): Words and Objections D. Reidel 1969

Davidson D (2): Essays on Actions and Events Oxford 1980

Davidson D (3): Inquiries into Truth and Interpretation Oxford 1984

Dummett M (1): Frege, Philosophy of Language: Duckworth 1970

Evans G (1): "Identity and Predication" J. Phil. 72 (1975)

Field H (1): "Theory Change and Indeterminacy of Reference" J. Phil 70 (1973)

Field H (2): "Quine and the Correspondence Theory" Phil. Rev. 83 (1974)

Field H (3): "Logic, Meaning and Conceptual Role" J. Phil 74 (1977)

Fodor JA (1): The Language of Thought Crowell 1975


Fodor JA (3): Representations Harvester 1981

Geach PT (1): Ercic Matters Blackwell 1972

Hintikka J and Suppes P (eds) (1) Information and Inference D. Reidel 1970

Hintikka J (2): Logic, Language Games and Information Oxford 1973


Kripke S (2) Naming and Necessity Blackwell 1980
Kripke S (3) Wittgenstein on Rules and Private Language Blackwell 1982
Loar B (1) Mind and Language Cambridge 1981
Lewis D (1) Convention MIT Press 1969
Lewis D (2) Philosophical Papers Vol I Oxford 1983
McGinn C (2) The Character of Mind Oxford 1982
McGinn C (3) The Subjective View Oxford 1983
Nagel T (1): "What is it like to be a bat?" in Nagel T Mortal Questions Cambridge 1979
Peacocke C (2) Sense and Content Oxford 1983
Perry J (1): "Frege on Demonstratives" Phil. Rev. 86 (1977)
Putnam H (1) Mind Language and Reality Cambridge 1975
Putnam H (2) Meaning and the Moral Sciences Routledge 1978
Putnam H (3) Reason Truth and History Cambridge 1981
Putnam H (4) Realism and Reason Cambridge 1983
Quine WVO (1) From a Logical Point of View Harper 1953
Quine WVO (2) Word and Object MIT Press 1960
Quine WVO (3) Ontological Relativity and other essays Columbia UP 1969
Quine WVO (4): "On the reasons for the indeterminacy of radical translation" J. Phil. 67 (1970)
Quine WVO (7) Theories and Things Harvard 1981
Saarinen E (ed) Game-Theoretical Semantics D. Reidel 1979
Searle J (2) Intentionality Cambridge 1983
Thalberg I (1) Perception, Emotion and Action Blackwell 1977
Wiggins D (1) Identity and Spatio-Temporal Continuity Blackwell 1967
Woodfield A (ed) (1) Thought and Object Oxford 1982
Wittgenstein L (1) Philosophical Investigations Blackwell 1963
Wright C (1) Wittgenstein on the Foundations of Mathematics Duckworth 1980
Wright C (2) Frege's Conception of Numbers as Objects Aberdeen UP 1983