THE UNIVERSITY OF OXFORD

Non-conceptual Psychological Explanation: Content and Computation

Ronald L. Chrisley
New College

Submitted for the degree of D.Phil

January 11, 1996

Contents

List of Figures

THE UNIVERSITY OF OXFORD

Non-conceptual Psychological Explanation:

Content and Computation

Submitted for the degree of D.Phil

January 11, 1996

Ronald L. Chrisley

New College

ABSTRACT

It is argued that a notion of non-conceptual content can help provide explanations of phenomena, such as the development from infant to adult cognition, that have so far eluded intentional explication. Chapter 2 introduces the Example, from the developmental psychology literature on object permanence in infancy, which serves as a benchmark throughout the thesis. After introducing the notion of a content-based explanation, the chapter argues that a content-based explanation of the Example is required, and yet objectual content-based explanations are inadequate. Chapter 3 investigates the notions of structure and conceptuality in detail, and uses the results of this investigation to propose a notion of non-conceptual content as content which is non-General, non-objectual, and individuated with respect to more than its role in judgement. Chapter 4 defends non-conceptual content from McDowell's recent criticism. Chapter 5 argues that conventional "that"-clause specifications of conceptual content will not work for non-conceptual content, and investigates various alternative means of specification, arguing that a successful candidate must employ computational concepts. That requirement, in addition to a general requirement for a naturalization of content ascriptions, motivates chapter 6's defense of the coherency of computational explanations in cognitive science, recently called into question by Putnam and Searle. With computation back on a safe footing, chapter 7 argues that the best computational vehicle to naturalize non-conceptual explanations is a sub-symbolic one. The particular case of a connectionist sub-symbolic system, the Connectionist Navigational Map, is described and used to illustrate the affinities between connectionist representations and non-conceptual content. Chapter 8 continues the use of this expository device, detailing an empirical investigation into the conditions under which the Connectionist Navigational Map can increase the conceptuality of its spatial contents, an explanatory approach that requires a notion of non-conceptual content.

ACKNOWLEDGEMENTS

The research for this thesis was conducted over several years, in several places, with support from several sources.

First, I wish to express my gratitude to my primary supervisor, Chris Peacocke, and to those who have also supervised my work at various stages: Kathy Wilkes, John Campbell, David Charles, Brian Cantwell Smith (who was a supervisor in fact if not in name), and especially Adrian Cussins.

Many of these ideas had their genesis while I was a part of a golden triangle, the vertices of which were Stanford University, Xerox Palo Alto Research Center (System Sciences Laboratory) and The Center for the Study of Language and Information. The latter two institutions provided the support which allowed me to supplement the first two years of my time at Oxford with visits to this triangle. While there, this thesis benefited from my friendships with many people, including: John Batali, Jim des Rivieres, Niklas Damiris, Matt Dixon, Vinod Goel, David Gurr, Tim Hawks, John Lamping, Erik McDermott, Geoffrey Nunberg, John Polito, Susan Stucky, Avery Wang, Helga Wild, John Woodfill, and Ramin Zabih.

At Oxford, I was fortunate enough to receive encouragement and advice from many people, including Bill Child, Mark Greenberg, Edward Harcourt, Reid Hoffman, Mike Martin, Gerard O'Brien, Lucy O'Brien, Ewan Paton, and Nigel Shardlow,

My research has been the recipient of support from institutions in two Scandinavian countries. In Finland, Paavo Pylkkänen and the Cognitive Science Research Group of The Finnish Academy, the University of Helsinki, and many individuals, including Teuvo Kohonen and Tere Vaden, were supportive and generous with time and resources. So also was the Connectionist Research Group of the University of Skövde, Sweden, including Mikael Bodén, Stig Emmanuelson, Jörgen Hansson, Dan Lundh, Jonas Mellin, and Lars Niklasson.

The School of Cognitive and Computing Sciences at the University of Sussex patiently provided the means by which the final stages of this thesis could be completed. Here I have benefited from discussions with, among others: Aysha Ahmed, Tom Beament, Maggie Boden, Jo Brook, George Butterworth, Terence Cain, Wendy Clements, Dave Cliff, Paul Davies, Zoltán Dienes, Matthew Elton, Richard Gaskin, Inman Harvey, Paul Hodgson, Bob Ives, Philip Jones, Ronald Lemmen, Michael Morris, Lindsay Parkin, Derek Parkinson, Murali Ramachandran, Ted Ruffman, Julie Rutkowska, Matthew Taylor, Chris Thornton, Mike Wheeler, Blay Whitby, Nicola Yuill, Tad Zawidski, and Jason Noble, who also patiently read and proofread a near-final draft, with very little notice.

Many, many other people have assisted me with their criticism and insight, including: Bill Brewer, Dave Chalmers, Paul Churchland, Andy Clark, Tim Crane, Daniel Dennett, Peter Dyan, Michael Dyer, Miranda Fricker, Mike Jordan, Brian Keeley, Dave Kirsh, Lisa Meeden, Ajit Narayanan, Stephen Neale, Kim Plunkett, Zenon Pylyshyn, Steven Schiffer, John Searle, Paul Smolensky, Dave Touretzky, Charles Travis, and Tim van Gelder.

Special thanks to Andy Holland, who implemented my design of the CNM in the C programming language, and ran the numerous simulations that yielded the results in chapter 8.

I benefited also from the comments and questions of the participants of several conferences and seminars, including: the 1990 Connectionist Models Summer School at UC San Diego, the 1992 Cognitive Science Summer School in Aix-en-Provence, the Exeter University Cognitive Science Group Seminar, the Gothenburg University Linguistics Seminar, David Rumelhart's Stanford Research Group Seminar, the University of Sussex Philosophy Society, the University of Sussex Cognitive Science Seminar, the University of Sussex COGS Research in Progress Seminar, the University of Sussex Developmental Psychology Research in Progress Seminar, the COGS Cognitive Philosophy Reading Group, the International Workshop on Android Epistemology, the First and Second Swedish Conferences on Connectionism, the International Workshop on Android Epistemology, the Wolfson Philosophy Society at the University of Oxford, and G. H. von Wright's Philosophy Research Seminar.

This research would not have been possible without further resources from the following institutions and individuals: San Francisco Branch of the English-Speaking Union, The Oxford Overseas Students Awards Scheme, Vice Chancellors of the UK Universities, Karen Chrisley, Adrian Cussins and Kathryn Moody.

To Annabel Weeden I offer my warmest thanks for the years of love and selfless support.

This thesis is dedicated to Kathryn, Karen, Kelsey and Hillary.

DECLARATION

I declare that this thesis is wholly my own work.

An earlier version of chapter 5 was published as "Non-conceptual Content and Robotics: Taking Embodiment Seriously," in Ford, K., Glymour, C. and Hayes, P. (eds.) *Android Epistemology*, Cambridge: AAAI/MIT Press, pp 141-166, 1995; and as University of Sussex COGS CSRP No. 246.

An earlier version of chapter 6 was published as "Why Everything Doesn't Realize Every Computation," *Minds & Machines* 4:4 (special issue on "What is Computation?"), pp 403-420, 1994. An even earlier version was published as "The Ontological Status of Computational States," *The European Review of Philosophy* 1:1; Stanford, CA: CSLI Publications, pp 55-75, 1994; and as University of Sussex COGS CSRP No. 247.

An earlier version of chapter 7 was published as "Connectionism, Cognitive Maps, and the Development of Objectivity," in Niklasson, L. and Boden, M. (eds.) *Connectionism in a Broad Perspective*, London: Ellis Horwood, pp 25-42, 1994; as "Connectionism, Cognitive Maps, and the Development of Objectivity," *Artificial Intelligence Review* 7, pp 329-354, 1993; and as University of Sussex COGS CSRP No. 259. Portions also appeared in "Non-conceptual Content and Parallel Distributed Processing: A Match Made In Cognitive Science Heaven?" in Trappl, R. (ed.) *Cybernetics and Systems Research '92, Vol. 2*, Singapore: World Scientific Publishing Corporation. pp 1335-1342, 1992. Other portions appeared in "Cognitive Map Construction and Use: A Parallel Distributed Processing Approach," in Touretzky, D., Elman, J., Hinton, G., and Sejnowski, T. (eds.) *Connectionist Models: Proceedings of the 1990 Summer School*, San Mateo, CA: Morgan Kaufman, pp 287-302, 1990.

An earlier version of chapter 8 was published as "Connectionist synthetic epistemology: Requirements for the development of objectivity," in Niklasson, L. and Boden, M.

(eds.) *Current Trends in Connectionism: Proceedings of the 1995 Swedish Conference on Connectionism*, Hillsdale, NJ: Lawrence Erlbaum, pp 283-309, 1995; and as University of Sussex COGS CSRP No. 353. pp 1-21.

Portions of chapters 2 and 3 have appeared in some of the above publications as well.

CHAPTER 1

Introduction

> I am eating breakfast. Hillary runs into the kitchen, looks at her watch, grimaces and groans, grabs the phone and begins dialling. Kelsey stumbles in, rubbing the sleep out of her eyes, and takes a seat. "Why is Hillary on the phone?" she yawns, reaching for the milk.
>
> "Because she believes she's late for work, and wants to let her boss know that she's on her way."
>
> Kelsey pauses, blinks, thinks a bit.
>
> "Don't you think we should tell her that today is Sunday?"
>
> "She'd probably think it's another practical joke on our part. She'll find out soon enough."

It's a fact that we understand each other in terms of what are called the propositional attitudes, attitudes (beliefs, desires, intentions, etc.) toward contents (roughly, propositions that have a correctness condition, such as *today is Sunday*; a fuller explanation of what content is is given in the next section). And the banality of the above example highlights the fact that such understanding, our *folk psychology*, is so pervasive as to be almost transparent to us.

Some philosophers (e.g., [Fodor, 1987]), focussing on the successes of folk psychology, and starting with the common-sense intuition that its explanations actually pick out the real causes of our behaviour, have proposed it as a foundation for a scientific understanding of mind. Perhaps some notions have to be tidied up, they concede, and perhaps some or even many of our folk generalizations involving these notions will be found to be incorrect, but the essence of the approach will remain intact. A scientific understanding of behaviour that does not employ the propositional attitudes or some variant of them will at best miss out on some illuminating patterns, laws, and regularities; and at worst fail to explain all but the simplest bodily movements. Indeed, orthodox cognitive scientists have adopted the common-sense notion of mental states as their theoretical starting point, even if they do not always actually employ them in the explanations they offer. Originally reacting to behaviourism, these cognitive scientists argue that the vast majority of human behavioural phenomena resist analysis by the traditional explanatory strategies of the natural sciences. Such phenomena are "intentional", exhibit "aboutness", are "other-directed", etc., and as such must be explained in normative terms. However, the chiding of the behaviourists has left at least this legacy: cognitive scientists believe that such intentional notions should be *naturalized*. That is, they feel that cognitive science should explicate both how it is that physical systems can be intentional, and how normativity can be a part of the natural order. Accordingly, traditional cognitive science offers a naturalistic means of explaining intentional phenomena, by appealing to representations and their attendant duality of vehicle (the physical manifestation of the representation, or, bluntly, syntax) and content (the significance of the representation, the way it presents the world as being; bluntly, semantics).

Other philosophers (e.g., [Stich, 1983] and [Churchland, 1981]), struck instead by the many *failings* of folk psychology, have come to the opposite conclusion: folk psychology is *not* a suitable foundation for a scientific understanding of so-called intentional phenomena. A proper, scientific understanding of behaviour, they argue, is non-intentional, or at least does not involve the propositional attitudes. They give several reasons for their conclusion, including the stagnant nature of folk psychology and its irreducibility to the natural sciences. But the most compelling reason that they give for questioning folk psychology as a foundation for cognitive science is its apparent failure to explain the putatively intentional behaviour of infants, animals, the pathological (i.e., mentally ill), intelligent artifacts, and anyone or anything else that differs substantially from us. And it is the intentionality of the content-based explanations at the heart of folk psychology that is to blame.

This thesis is concerned with some aspects of a fundamental and far-reaching question that arises from this debate: can one have a content-involving science of the behaviour of not just adult humans, but of other apparently intentional systems such as infants, animals, artifacts, etc.? However, many of the particularities of the debate concerning the role of folk psychology in science are irrelevant for the purposes of this thesis. For example, I am not here concerned with the fate of folk psychology *per se*; rather, I am concerned with the possibility of *any* scientific explanation of behaviour that involves the notion of content, whether or not it can be seen as an extension of our folk psychology.[1] The main claim of the thesis is that a notion of *non-conceptual content* greatly increases the prospects for a contentful psychology.

Therefore, chapter 2 starts with an examination of propositional attitude explanation on the one hand, and the intentional phenomena that appear to slip through its nomological net (what I call the *Recalcitrant Phenomena*) on the other. For the sake of generality, propositional attitude explanation is generalized to content-based explanation (to accommodate, for example, explanations involving the contents of sub-personal states). Then the Recalcitrant Phenomena, and the difficulties they cause for content-based explanation, are described.

This is achieved by focussing on an Example: a particular case of the behaviour which traditional accounts have been unable to explain satisfactorily. The Example used is the pre-“object permanence” behaviour of infants. To establish the necessity of an alternative account for this Example of a Recalcitrant Phenomenon, it is shown that traditional accounts (non-contentful on the one hand, or conceptual on the other) are inadequate. Specifically, it is shown how accounts employing only conceptual content cannot succeed, because they necessarily involve systematicity and generality which is inappropriate for the Example.

Thus, one cannot give traditional content-based explanations of the Example; the set of truly Recalcitrant Phenomena is non-empty. If one further assumes that content-based explanations are the only ones available for scientifically explaining intentional phenomena, then a dilemma is forced: either the phenomena covered by the Example are, despite appearances, non-intentional; or they cannot be scientifically explained at all. Some might (as I do) find neither of these options acceptable; what is one to do?

I reject the dilemma by rejecting one (or both) of the two premises that forced it. The premises are:

1. that one cannot give traditional content-based explanations of the Recalcitrant Phe-

[1] For example, one might think that a psychology that did not appeal to *concepts* would not be an extension of folk psychology; yet as will be explained, I believe that the success of a purely *non-conceptual* psychology could vindicate the content-based approach.

nomena; and

2. that traditional content-based explanations are the only means of scientifically explaining intentional phenomena *qua* intentional.

This rejection is effected by sketching out an alternative intentional account that *can* accommodate the Recalcitrant Phenomena. Which premise it is that one will see me as rejecting depends on how strong a reading one gives “traditional”. On a narrow reading of “traditional”, the approach I am taking is sufficiently radical to be considered a rejection of the traditional method's monopoly on scientific intentional explanation (2). On a more catholic understanding of what traditional content-based explanation comprises, the approach I advocate here is an extension of the basic notions of content and representation; it was the assumption that traditional explanations could not account for the Recalcitrant Phenomena (1) that was mistaken. Partly to avoid this uneasy ambiguity, I focus on one aspect of traditional accounts: their conceptuality. The two premises can then be rephrased with “conceptual” instead of “traditional”. I will make it clear exactly what is meant by a conceptual account, and thus the ambiguity of “traditional” will be eliminated. It can then be seen that it is the second premise that I am denying: there are intentional explanations of phenomena that do not employ exclusively *conceptual* content-based explanations.

Accordingly, chapter 3 details the notion which is the centerpiece of the alternative account that will admit the Recalcitrant Phenomena: non-conceptual content. I join others [Crane, 1992, Cussins, 1990, Davies, 1990, Evans, 1982, Haugeland, 1991, Peacocke, 1993] in arguing that a distinction should be made between conceptual and non-conceptual contents.[2] For several reasons, much of the work in cognitive science has concentrated by default on the case of conceptual content, but there is reason to believe that understanding non-conceptual content is essential to understanding (and therefore to designing) intentional systems in general [Cussins, 1990]. A definition of non-conceptual content that has enjoyed some popularity is: content the entertaining of which does not require a subject to possess the concepts used to individuate that content (e.g., [Peacocke, 1986, p 17] and

[2] I should point out that none of the cited authors characterize non-conceptual content in exactly the same way, nor does my notion agree with any one of theirs. These differences will be examined in chapter 4.

[Cussins, 1990, p 382 ff]). Although this is a useful definition for explaining some aspects of some of the Recalcitrant Phenomena (particularly those concerning contents involved in perception and action), a preferable definition for my purposes is: non-conceptual content is content the constituents of which are not all concepts. This definition is of use only if one has a clear understanding of what constituents of contents are, and what conditions must be met in order for a constituent to be a concept. I provide an account of these issues in chapter 3: concepts are content constituents that are individuated with respect to their role in judgement alone.

It is shown there that a consequence of this understanding of concepts is that the Generality Constraint holds [Evans, 1982, page 104]. That is, for any concepts (ways of thinking of properties) F and G, and any concepts (ways of thinking of objects) a and b, if a subject knows what it would be for a to be F and for b to be G, then it must know what it would be for a to be G and for b to F. It is also shown that the corresponding constraint for non-conceptual contents need not hold. It is this fact which allows non-conceptual content based explanations to address the Recalcitrant Phenomena in general, and the Example in particular, which are conceptually inexplicable. The notion of non-conceptual content being put forward here is contrasted with the notions of other writers: e.g., non-conceptual content can be the object of attitudes such as belief; and it can have a determinate truth value.

Building on this analysis, chapter 3 closes by showing that non-conceptual content is more than a mere formal possibility; its lack of Generality allows it to provide explanations of the Example, a phenomenon which cannot adequately be explained by more conventional means.

McDowell [McDowell, 1994b] has recently criticized the notion of non-conceptual content, arguing from a neo-Kantian position that content cannot be reason-providing if it is outside the conceptual sphere. Yet it must be able to provide reasons if it is to explain the Recalcitrant Phenomena under consideration. Chapter 4 analyses and gives reasons for rejecting McDowell's objections.

Although chapter 3 gives an account of how non-conceptual content should be individuated, the extremely theoretical nature of this individuation scheme renders it fundamentally inappropriate for the explanation of particular intentional phenomena. By way of analogy, consider how infeasible it would be to specify *conceptual* contents in our explanations of each other by enumerating their logical relations to other contents, their acceptance conditions, etc., rather than by using standard "that"-clause specifications. However, chapter 5 argues that "that"-clause specifications cannot be used for scientific applications of non-conceptual content. Therefore, some alternative means of specification are suggested and evaluated. Of the authors that have advocated the notion of non-conceptual content, only Peacocke ([Peacocke, 1989], [Peacocke, 1992], [Peacocke, 1993]) has offered detailed ways in which we might specify them. However, his focus on a particular subset of non-conceptual content (perceptual contents), along with his different vision of the role that non-conceptual contents play in explanation, means that his notions of *scenarios* and *proto-propositions* cannot perform the specification task demanded by the kind of non-conceptual content which is of use in addressing the Example in particular, or much of the Recalcitrant Phenomena in general. Accordingly, chapter 5 offers some means of specification that will work for such contents.

In the cognitive scientific explanatory scheme, contents (even non-conceptual ones) need vehicles. Naturalism requires that we show how our intentional and non-intentional characterizations of an organism cohere; the hypothesis of cognitive science is that we should do this by finding a *computational* characterization which marches in step (a relationship which is weaker than reduction) with our intentional account [Cussins, 1987, Dennett, 1987, Fodor, 1985]. Chapter 6 clarifies this computational means of naturalizing intentionality. In order to clear the way for developing the connections between non-conceptual content and computation, chapter 6 also gives reasons for rejecting some recent criticisms of the role computation can play in a scientific understanding of the mind. Specifically, both Putnam [Putnam, 1988] and Searle ([Searle, 1990] and [Searle, 1992]) have presented arguments for the claim that computational states are universally realisable, in the sense that we could interpret any physical system as instantiating any computational characterization. They both argue that this has dire consequences for the computational view of the brain and mind that is a working hypothesis in cognitive science. But whereas Searle admits that the threat of universal realisability could be avoided if our notion of computation involved causal and counterfactual notions (implying that these are lacking at present), Putnam thinks that the universality, and hence vacuousness, of the notion of computation remains, even if one requires computational state transitions to be causal. Chapter 6 analyses Putnam's argument and finds it inadequate, because it employs a notion of causation that is too weak. It argues that if one were to augment formal notions of computation with constraints involving embeddedness and full causality, one would arrive at a notion that directly addresses the worry of universality.

With computation back on a firm footing, chapter 7 explores the constraints that non-conceptual content places on representational vehicles. In addition to showing how a notion of non-conceptual content can explain sub-symbolic computational systems, chapter 7 explores how sub-symbolic computation can naturalize non-conceptual ascriptions and explanations. It is argued that a computational architecture appropriate for naturalizing non-conceptual content will differ substantially from classical, purely symbolic notions of computation. In particular, chapter 8 shows how the gradedness of the systematicity of non-conceptual contents is accommodated well by the dynamics of learning in the Connectionist Navigational Map. The chapter again appeals to the learning capacities of the Connectionist Navigational Map as an expository device. The process of *constructing* a map, moving from a state of little or no knowledge of the topological and featural properties of places to a state in which a terrain can be navigated successfully, is seen as a paradigmatic case for the need of an analysis in terms of a transition from non-conceptual content to (relatively) conceptual content.

CHAPTER 2

Content-based Explanation and the Recalcitrant Phenomena

Before providing an Example of the Recalcitrant Phenomena (the phenomena which cause difficulties for traditional content-based explanation, but which can be accounted for with a more general notion of content), I say more about what content-based explanation is, starting with an explanation of what content is.

2.1 Content

Traditionally, content is defined as "that which is expressed by an utterance or sentence" [Blackburn, 1994, p 79]. However, the notion of content with which I am concerned is *mental* content, and has its primary application, not concerning language, but the mental states of experiencing subjects – for example, the objects of the propositional attitudes. I do not want to *define* such objects to be linguistically expressible; indeed, in chapter 5 I argue that there are contents which cannot be so expressed (on a strict understanding of "expressible"). Since I am focussing on propositional attitude explanations (as explained later in this chapter), it is tempting for me to adopt the corresponding definition of mental content as "that which is believed (desired, intended, etc.) by an experiencing subject". But I will resist this temptation, since I recognize the existence of classes of content, which, unlike the class of non-conceptual content which I will be putting forward, cannot be the objects of the propositional attitudes.[1] Rather, the fundamental notion of content is this: the way the world is presented, in experience, to a subject.

Some comments on this notion of content are in order. Since it is only the *fundamental* notion of content from which others are derived, it does not rule out notions of content that are personal, yet non-experiential; nor does it rule out sub-personal or sub-organismal content (for the distinction between sub-personal and sub-organismal content, see chapter 4). But it does stipulate that these notions of content must derive from the fundamental notion of content, which involves an experiencing subject.

The fundamental notion relies on two key elements: the idea of the world, or an aspect of the world, being presented in experience; and the idea of a way in which that presentation is made. Correspondingly, each content has two essential features: characteristic referential properties (a semantic value), which determine which aspect of the world is being presented, and a characteristic pattern of psychological significance, which determines the way in which that aspect of the world is being presented. The fact that both of these factors are constitutive of content can be seen another way. Experiences involve, in general, different aspects of the world. So what is characteristic of a content involves, at least, its semantic value. But two experiences may involve the same aspect of the world (their contents may have the same semantic value), and yet be distinct, because they present that same aspect in different ways. Thus, the pattern of psychological significance, together with the semantic value, is characteristic of each content.

There is another way of speaking, in which content is sopken of in the way one speaks of Fregean sense. Thus, content itself does not strictly have a semantic value, but rather the term "content" refers to the psychological significance of something else (say a thought, or an expression), which does have a semantic value. But the differences in these two views can be ignored for the purposes at hand. For example, one could consider "the semantic value of content C" merely to be shorthand for "the semantic value of whatever it is that has C as its content".

If content is to be understood as the way something is presented in experience, more

[1] I have in mind the 'horizontal' notion of non-conceptual content recently discussed in the philosophy of perception literature; see chapter 3.

must be said about the notion of things being presented in experience in different ways. The foregoing already started on this task, by understanding the ways of being presented in experience that are constitutive of a content C as being the characteristic pattern of psychological significance of the mental state that has C as its content. Thus, attention is shifted to the notion of psychological significance.

Psychological significance is a generalization of the notion of cognitive significance. To understand this remark, it will be necessary to look at the historical origins of the modern notion of mental content.

Frege noted that two terms with the same semantic value could nevertheless differ in other ways, ways that fall into two kinds. First, they may differ in ways that, for Frege, do not result in a difference of content. For example, they may differ in their *tone*. Even if we suppose that 'horse' and 'steed' have the same semantic value, they nevertheless differ in tone; yet Frege took this to not be a difference in their contributions to thoughts. On the other hand, 'Hesperus' and 'Phosphorus' share the same semantic value, yet they have a difference which *is* reflected in their content. We can see that their contents differ, because we can see that there are sentences, e.g. identity statements, that differ only with respect to these terms, that themselves differ in informativeness. For Frege, difference in tone alone does not yield a difference in content; but difference in informativeness does.

From this alone one can derive a sufficiency criterion for distinctness of content (such as Frege's criterion of difference, see also [Peacocke, 1993, p 2]), but more is needed for a theory of content. What deeper truth is at work when one descries a difference of content based on the informativeness of an identity statement? Traditionally, this has been answered with a notion of cognitive significance: a content is individuated by its role in judgement. A difference in the informativeness of identity statements implies a difference of content because the former is a special case of difference of role in judgement, or cognitive significance. A difference in tone only is not a difference in content, because it is not a difference of role in judgement (cognitive significance).

Let us look ahead for a moment, to the the larger aims of this work. As stated in the introduction, it is my goal to provide explanations for the Recalcitrant Phenomena in general, and the Example in particular, by generalizing our notion of content. For reasons which are given in chapter 3, this should not be achieved via a generalization concerning semantic values. Rather, it should be done by generalizing the other part of the traditional standard against which content is individuated – cognitive significance – to a more inclusive standard, that of psychological significance. But how are we to understand psychological significance? If individuation according to cognitive significance is individuation with respect to judgement, must individuation according to psychological significance be individuation with respect to non-normative aspects of the subject, such as mood, emotion, realizing substrate, etc.? If so, this would indeed be a severe departure from Fregean orthodoxy; it is unclear what common ground, if any, would exist between the traditional view of content and this general view. More to the point: these differences do not seem to involve differences related to truth. But a notion of content that does not have truth as its unifying pole is not a notion of content at all; to ask one to "generalize" the notion of content in this way is to ask one to change the subject. Rather, the generalization from cognitive significance to psychological significance I propose is one that involves more than judgement: non-conceptual contents are not individuated with respect to their *acceptance* conditions alone, but with respect to their *entertainment* conditions in general. Individuation with respect to judgement alone focusses on the questions: On what grounds should a subject accept the content (judge it to be true)? What should the subject do if the subject does accept the content? Individuation with respect to entertainability asks these as well, but also asks different, but no less normative, questions: Under what conditions should a subject be able to enjoy states that have the content? What should the subject do if the content is entertained in a particular way? And if the subject is capable of entertaining the content, what other contents (including even those that are individuated with respect to judgement alone) should be entertainable if the subject has further experiences of a certain sort? These issues come to the fore again in chapter 3, particularly in sections 3.5 and 3.7; there it is argued that the properties of conceptual content derive from individuating content with respect to cognitive significance alone, and that the possibility of non-conceptual content is created when one individuates content according to psychological significance.[2]

[2] At one point, Cussins also chooses to understand non-conceptual content as content "whose identity

A final terminological point concerning content: those contents which have either truth or falsity[3] as their semantic value, contents which are evaluable, I will call *whole* contents; those that have a semantic value that must be combined with other semantic values to yield either truth or falsity I will call *partial* contents.[4]

2.2 Content-based explanation

I am defending the possibility of a particular kind of psychological explanation: content-based psychological explanation [McGinn, 1989, p 121]. My main contention is that recent criticisms of the possibility of content-based explanation (CBE) in a scientific psychology are based on a narrow view of the possible forms that CBE might take.

The most common form of CBE understands psychological states in terms of attitudes (belief, desire, knowledge, intention, etc.) toward contents (that there is a door ahead, that $2 + 2 = 4$, etc.); such attitude/content pairs are then used to explain other attitudes or action. For example, one might explain why an agent opened a door (i.e., one might show that the agent's opening of the door wasn't mere accident; if circumstances had changed slightly – if the door were one foot over to the right, say – the agent would have exhibited (different) behaviour that nevertheless opened the door) by claiming that the agent *intended to open the door*; one could explain the possession of this intention as being the result of the agent's *desire to be in the next room* and its *belief that opening the door*

conditions are fixed, not just by its constitutive connections to judgement, but by its constitutive connections to perception, action and judgement" [Cussins, 1990, p 389]. However, he develops this position in a way different from mine; I make clear some of the differences in section 3.8.1 of chapter 3.

[3] Others would insist that a generic notion of content should recognize true/false as just one possible norm of correctness. For example, Cussins [Cussins, 1990] has argued that non-conceptual content involves norms of correctness other than true/false (e.g., "appropriate/inappropriate", "adaptive/non-adaptive", etc.). I review his argument in chapter 3, section 3.8.1, and reject its conclusion: I do *not* wish to argue for a norm of correctness for non-conceptual content that is anything other than truth. Again, introducing other norms of correctness might be *another* way of generalizing content and therefore generalizing content-based explanation, but it is not the way that I will do so.

[4] This distinction between whole and partial contents is usually made by using the terms "propositional" and "sub-propositional", but I will avoid this terminology, since it often carries connotations and implications that I do not wish to incur.

will help one get into the next room.

Many criticisms of CBE attack the particular properties of the propositional attitudes, yet one might find a role for content in scientific psychological explanation that does not employ the notion of propositional attitudes. This would be one way of establishing my contention, but I do not pursue that possibility. Rather, I consider explanations which do take the traditional propositional attitude form, and instead argue for a more general notion of content.[5] Specifically, the principal criticisms of CBE (e.g., [Stich, 1983] and [Churchland, 1981]) have assumed that all content is conceptual content, yet I argue in chapter 3 that there is a class of contents which is *non-conceptual*, and that CBEs which employ this kind of content can avoid the limitations that might have made the prospects for conceptual CBE seem so bleak. Not only will this come as some relief to the friends of CBE, but it will also remove the pressure for them to take the desperate positions (nativism, immunity of CBE to elimination through empirical advances, et al.) that they have felt one must take in order to defend CBE against the eliminativist onslaught.

It might be instructive to examine how the essential features of content play a role in CBE. The role of the cognitive significance of a content (which is part of its psychological significance) is particularly crucial. For example, the following is explanatory: "Edward is going to the stables because a) he wants a steed and b) he believes that he can get a horse if he goes to the stables". For the present explicatory purposes, this explanation can be re-worded as: "Edward is going to the stables because a) he wants it to be the case that Edward has a steed and b) he believes that going to the stables will help bring it about that Edward has a horse". This invokes an implicit psychological generalization: "if x desires that it be the case that P, and x believes that doing A will help bring it about that P, then, ceteris paribus, x will try to do A". This generalization depends on the identity of the contents that stand in for each occurrence of the schema variable P. This identity condition is satisfied in the case of the explanation of Edward, because despite their different expressions, the contents of the attitudes indicated by "that Edward has

[5]This self-imposed restriction to the attitudes does not, however, limit the generality my results. There is no reason to believe that my findings will not be of use to those who *do* choose to investigate the possibility of CBEs which do not invoke the attitudes. In particular, one might find the notion of non-conceptual content useful, even essential, for sub-personal CBEs which do not invoke the attitudes.

a steed" and "that Edward has a horse" are the same. That is, because the cognitive roles of those two contents are the same, the psychological generalization can be invoked to explain the action. Note that this is not the case with an explanation like "Edward is going to Dummett's lecture because a) he wants to meet the greatest living philosopher and b) he believes that going to Dummett's lecture will help bring it about that he meets Dummett". The content "that Edward has met Dummett" does not have the same cognitive significance as the content "that Edward has met the greatest living philosopher", therefore any invocation of the psychological generalization is illicit.

Similarly, some CBE depends on identity of semantic value. The precise nature of this dependence is controversial, but let me arbitrarily assume a particular theoretical stance in order to provide an *example* of how CBE can depend on semantic value (for there are some who think that cognitive significance exhausts the role of content in the explanation of action; see [McGinn, 1982]). Consider the putative explanation: "Edward is going to the lecture because a) he wants to meet you and b) he believes that going to the lecture will help bring it about that he meets you". There are some accounts of demonstratives that say that the cognitive (or even psychological) significance of "you" is the same, no matter to whom it is referring. The content of "you", on this account, would be some mapping from different contexts to different referents, different semantic values. Assume such an account is correct. Then consider the putative explanation in the case when the two occurrences of "you" refer to two different people. Ex hypothesi ("you" always has the same cognitive significance), there is identity of cognitive significance, but clearly invocation of the psychological generalization should not be permitted (Edward's belief that going to the lecture will help him meet you, Brian, does not explain his action when taken together with his desire to meet you, Aaron). A further restriction on the application of the generalization needs to be provided, one that states that in some cases, not only must the contents share cognitive significance, but they must also share semantic value or reference. Thus, there are cases in which the conditions for correct application of a psychological generalization cannot be exhaustively stated in terms of cognitive (or even psychological) significance alone, but must under some conditions invoke (identity of) semantic value.

2.3 The Recalcitrant Phenomena

Although I do not attempt to provide a general theory of what the Recalcitrant Phenomena have in common such that they are inexplicable by orthodox content-based means, a little can be said about the kinds of intentional phenomena which cannot be explained via traditional means, but which, I claim, can be explained by a more general notion of content. First, there are intentional phenomena that are different from how we typically think of adult human cognition, in that they seem to be *pre-objective*. The mental lives and behaviour of infants and animals, for example, seem to have elements in which parts of the independently-existing world are being thought of (reference), but not *as* parts of an independently-existing world (sense). Next, there are the intentional phenomena involving real-time perception and action integration. And there are the dynamic aspects of intentional systems, be they evolution, morphogenesis, development, learning, or a change of conceptualization.

I am not defining the recalcitrant phenomena to be all those intentional phenomena for which traditional content-based explanations do not work. There may be many such phenomena which are unredeemable by the alternative account I am giving. Rather, I use the term "Recalcitrant Phenomena" to refer to those intentional phenomena which are inexplicable for traditional content-based accounts, but are explicable by non-conceptual content based-explanations. Of course, such a definition does not guarantee that there actually *are* any such phenomena. The purpose of looking at the Example is to establish that there are indeed such phenomena which require non-conceptual accounts.

Having an Example of the Recalcitrant Phenomena to hand will serve useful functions.

1. It will allow it to be made clear how it can be that there are phenomena for which a non-intentional account is not sufficient, and yet which cannot be fully explained using conventional notions of information or content;

2. It will suggest what it is about the phenomena that makes them so difficult to capture in terms of conceptual content, and what it is about the conventional accounts that prevents them from succeeding;

3. It will serve as a benchmark against which to test the utility of the notion of non-conceptual content which I will propose.

Concerning the third function: I could instead take the philosophical hard line, conducting my investigation into non-conceptual content in a purely a priori manner, leaving it to psychologists or cognitive scientists to determine if the concept has any use in the sciences of mind and behaviour. But this would be completely at odds with the main purpose of my investigation, which is to provide new conceptual tools for those wishing to provide empirical explanations of intentional systems, tools that will widen the scope of intentional phenomena that can be scientifically understood. True, the development of these tools will be a primarily a priori endeavour; but such conceptual work will be done with an eye towards a final reckoning: the utility of those tools for empirical explanation. Nevertheless, the prospects for the notion of non-conceptual content should not be limited to its success in accounting for the examples on which I choose to focus.

The Example focusses on the pre-object permanence behaviour of infants. Further phenomena, concerning sub-symbolic computation in general, and the learning of cognitive maps in particular, are covered in chapters 7 and 8. These phenomena also provide a motivation for non-conceptual content, but it is easier to see why it is that they do so after non-conceptual content has been thoroughly introduced in chapter 3, and after the architecture of the Connectionist Navigational Map is detailed (in the first part of chapter 7). By contrast, the implications of the Example proper can be clearly stated now, without, e.g., further clarification of non-conceptual content.

The data that suggest a lack of a conception of object permanence in infants provide a clear illustration of how the notion of the sub-, pre-, or non-conceptual can fill a long-empty explanatory gap in our understanding of human cognition. It is a good Example not only because the mental operations of the infant at a given point prior to full conceptuality are best understood in terms of non-conceptual content; but also because we can make sense of transitions which *increase the conceptuality of* an infant's interactions with its environment, and this dynamic also requires a non-conceptual understanding.[6] One minor

[6]The notion of degrees of conceptuality, how it is suggested by the Example, and why it requires a notion of non-conceptual content are discussed in chapter 3, section 3.7.

drawback of the Example is that some might mistakenly infer from it that non-conceptual content is not of use once the infant has developed, i.e., that non-conceptual content is not crucial for explaining adult human behaviour.

After stating the Example, I show why a purely non-intentional account of the phenomenon is not possible. Then, I show why a traditional, conceptual, content-based account is not satisfactory, simultaneously hinting at how a non-conceptual account might fare better. In the next chapter, I explain in detail the notion of non-conceptual content I am using. Then, with a clear understanding of non-conceptual content, it is possible to return to the Example (in section 3.7) and quickly show how a non-conceptual account can avoid the limitations that rule out the conceptual account.

2.4 The Example: Infants and object-(im)permanence

> **The Example:** At an early age, children have intentional interactions with objects that they are currently perceiving, yet they do not think of these objects *as* independently-existing objects. This way of thinking of objects is only acquired later, through experience, and in stages.

Later, when showing why conceptual content based explanations fail for the Example, I present the data from which the Example is drawn, following closely the presentation in Paul Harris' recent review article "Object Permanence in Infancy" [Harris, 1989]. But first, I show why the Example warrants contentful explanation.

2.4.1 Why a contentful account is warranted: Perspectival sensitivity

The Example is characterized in such a way that it guarantees the applicability of an intentional account: "children have intentional interactions with objects that they are currently perceiving". But what is it about the behaviour of infants that warrants describing their interactions as intentional? And what is it that rules out a purely mechanistic account of these infants? Of course, a non-intentional account need not be ruled out in order for an intentional account to be warranted (there may be correct accounts of a phenomenon in more than one kind of discourse), but there should be some aspect of the data that a non-intentional explanation fails to account for. In the Example, the relevant data are:

> **Tracking data:** While playing with a toy under certain conditions, an infant will track the toy if it is moved in view of the infant; the infant will reach for the toy if it is in view.

Someone employing the notion of the infant trying to track the object might be very successful at predicting and explaining the infant's arm motion, direction of gaze, etc. in a way that would far outstrip the explanations and predictions given by, say, a non-intentional neuroscience. Consider the prediction: "the infant will, ceteris paribus, point in the believed ego-centric direction of the toy; the infant will, ceteris paribus, believe the toy to be located in ego-centric space where it actually is located in that space". This simple, intentionality-invoking rule will be successful in a wide range of cases, a range which could only be covered by multitudinous and complex neurophysiological conditionals. Thus, a non-intentional account would be inadequate.

So also would a non-contentful account be inadequate. Peacocke has considered the necessary and sufficient conditions for ascribing content in this case, which he calls the Basic Case: "the question of what it is for a subject to have attitudes about places and perceptible objects in his immediate environment" [Peacocke, 1983, p 57]. Most of his discussion of the Basic Case is concerned with the necessary conditions for the ascription of such contents. He argues that an informative necessary requirement for such contents is the condition of *perspectival sensitivity*:

> [A] simplified general statement of the requirement of perspectival sensitivity would be this: if the subject moves from one place to another, his intentional web must be recentred on the place determined in normal circumstances by the change in sensational properties[7] of his experiences. [Peacocke, 1983, p 69]

In brief, an intentional web is a set of ways that objects are given to a subject in perception. These ways are individuated by indicating the actual direction in which a subject would move if the subject acted upon an intention to move towards the object. Recentring the web is updating the ways in which those objects are presented in perception, so that even though, say, the subject has moved to a new position, the intentions to move

[7] No doubt Peacocke would now wish to re-formulate this in terms of some form of content, such as scenario content, rather than in terms of sensational properties of experience; this does not affect the point being made.

related to those ways of thinking would still cause movement towards the actual position of the object.

An example of a subject failing to meet the condition of perspectival sensitivity is this: suppose a subject is facing north, and is aware of a desirable object directly on his right, to the east. The way that the object is being thought of entails that if the subject wanted to move towards the object, he would go east. So far so good. However, suppose that the subject moves north, yet does not recentre his intentional web. That is, the object (now to the southeast) is still being thought of in a way that, if the subject were to want to move toward it, would cause the subject to move directly east. That would be a case, then, of a failure to be perspectivally sensitive, and without further evidence would be enough to undermine all putative talk about ways of thinking of, intentions to move toward, etc. the object in question.

The point is that the infant in the tracking data for the Example is not like the insensitive case; the tracking infant is perspectivally sensitive. In tracking an object, the infant is updating its intentions to look, point, reach, etc. in just the sense that recentring the intentional web requires. Of course, in the data we are considering it is the object that is moving, not the infant, but the application of the notion of perspectival sensitivity is symmetric.

Although Peacocke refers to perspectival sensitivity as a necessary, and not sufficient condition, he means only that it is necessary and not sufficient for full-fledged *judgement* of spatial contents; he agrees that it is sufficient for content simpliciter:

> Perspectival sensitivity is necessary for the ascription of content, but it may reasonably be questioned whether it is sufficient for thought in the way that human beings are capable of thinking. The condition of perspectival sensitivity could be fulfilled by a being whose cognitive psychology is completely describable by the fact that he updates a a map of his environment on the basis of his experiences, and this map interacts with his desires to produce action. There is indeed genuine content here, but not judgement: in particular, this being never withholds assent from the proposition which is the content of an experience, in order to weigh up the evidence or to explain away recalcitrant experiences. Recalcitrant experiences are first taken at face value, and later ignored if they cause errors, but they are not reflected upon. It would be dogmatic to claim that the word 'belief' in English excludes the states of such a being, but there is clearly a real distinction between the case of this being and our own thinking. Hume's conception of belief as determined by brute custom is applicable to this being in a way it is inapplicable to us. But such a being

> does register facts about its environment, and has psychological states with content [Peacocke, 1983, p 77-78].

By separating out the the presence of content from the case of content coming under the purview of judgement, possible objections to a contentful account of the Example can be seen to be objections only to adult, judgement-involving contentful accounts. Without such an objection, any success of a contentful account in explaining and predicting the infant's behaviour, a success which cannot be reproduced at the non-intentional level, is enough to justify content-based explanations of the Example.

There is an example that might make one doubt the claim that the condition of perspectival sensitivity, despite the assurances of the author of *Sense & Content*, really does demand a content-based account. The example is this: consider a simple toy car, with a compass. Suppose that the motor and wheels are hooked up to the compass so the car has a disposition to move toward the Earth's north magnetic pole. It will be the case that no matter where one moves the car, the disposition of the car to move will be updated accordingly. The car will recentre its intentional web. And it will do this in the light of the changes in the sensational properties of its experiences: the change in the direction the compass is indicating. Thus, the simple toy car with compass meets Peacocke's condition of perspectival sensitivity. And yet few would want to admit that there is a case of content here. The reluctance to ascribe content is not due to an anthropocentric chauvinism, but rather an aversion to panpsychism: if something as simple as this toy car and compass has content, then practically every physical system will have content and intentionality.

One way of blocking the conclusion that the toy car is perspectivally sensitive is to take a hard line on experience. As stated above, the definition of perspectival sensitivity requires that the recentring of the intentional web be dependent upon "the change in sensational properties of [the agent's] experiences". If the toy car has no experiences, which seems likely, then any putative recentring of its intentional web, if it can still be said to have one, is occurring on the basis of some mere causal link, not on the basis of changes in experience, as Peacocke requires. Thus one can safely deny content to the toy car, while simultaneously ascribing content, as before, to the infant in the Example. For in the cases involving the infant, we do have prima facie reason to believe that there is experience present, and that it is on the basis of changes in this experience that the recentring of the infant's intentional web is taking place.

Although taking this hard line is tempting, since it permits me to re-assert that the Example does require a content-based account, I will not avail myself of it. On the hard line, it is in virtue of an account of experience that one can have grounds for attributing particular experiences to some systems, and it is in virtue of those attributed experiences, plus other facts, that a system has content. Thus, the hard line assumes that an account of experience is already available, prior to an account of content. This arrangement does not seem plausible. Rather, it seems more plausible that an account of content will be developed concomitant with (if not prior to), rather than rely heavily upon, an account of experience.

Thus, I prefer giving another response to the panpsychic threat posed by the toy car and compass. The fact is, the toy car, as described, does not actually meet the conditions of perspectival sensitivity. In Peacocke's original example of an agent recentring its intentional web, the agent has dispositions to move toward locations, but these dispositions were of a sort that could be pre-empted by other dispositions of the agent itself. That is, the agent might have a disposition to move to some location off to the right, but a disposition to move straight ahead could over-ride the right-directed intention. This would result not only in the agent moving straight ahead, but also in it updating its right-directed disposition to a right-and-behind-directed disposition. This kind of persisting disposition in the face of other, overriding dispositions originating from the agent itself is not present in the simple toy car plus compass example. There may be times when the toy car is unable to exercise its disposition to move toward the magnetic pole: the toy car may be buffeted by external forces, such as a child picking it up and placing it elsewhere. Yes, in these conditions it can recentre its intentional web appropriately: once the external restraining condition is removed, the toy car will once again move toward the magnetic pole. But what the simple toy cannot do is fail to exercise its disposition to move toward the magnetic pole because of *another of its own intentional dispositions* overriding the one directed toward the magnetic pole. Therefore one can withhold ascribing perspectival sensitivity, and with it content, to the toy car.

But we are once again taking Peacocke's condition of perspectival sensitivity on faith. Why should one think that the ability to recentre one's intentional web in the face of other internal overriding dispositions is sufficient for content?

In a real sense, this question cannot be answered here. But it is important to recall the dialectical role of the consideration of perspectival sensitivity, to see why not being able to answer this question fully is not damning for this project. Perspectival sensitivity was brought up as a way to add credence to the claim that the Example requires a content-based account. But it would be a mistake, given the objectives of this project, to think that there should be, *prior* to a new account of content that the project aims to supply, a proof that the Example requires a content-based account. Rather, the conclusive way to establish that an empirical phenomenon warrants a content-based account is to provide such an account. It is my position that the best contentful account of the Example is non-conceptual. Thus, it is reasonable for me to maintain that conclusive establishment that the Example warrants a contentful account can only be provided after the notion of non-conceptual content has been explained in detail. Then the success of a non-conceptual account can *justify* deeming the Example to be content-involving. Yet the Example is meant to motivate *raising the subject* of non-conceptual content; it is meant to convince one of the value of inquiry into what non-conceptual content *is*, before that inquiry has properly begun. Therefore, at this point, one can at most hope to establish the *plausibility* of a content-based account for the Example; asking for conclusive evidence that it is contentful would be, in that it is premature, circular.

Thus, perspectival sensitivity is invoked to show the plausibility of a content-based account of the Example. This plausibility is increased further by refuting a second objection to the claim that perspectival sensitivity is sufficient for content. The second objection follows from the first: could one not design a toy car that had the overriding dispositions that the toy car in the first objection lacked? It seems one could. The toy car would have the compass and dispositions as before, but would also have a video camera attached, and be designed so that it has a disposition to move toward the largest red thing in its camera's visual field. This disposition would be updated through an appropriate recentring of the toy car's intentional web. Furthermore, the toy car could be designed so that the red-directed dispositions always override the magnetic dispositions. Now suppose the toy car is moving toward a red ball on the other side of a depression to the East. As long as the red ball is in the camera's visual field, the dispositions to move toward the ball will be updated and will control the car's movement. But the car's disposition to move towards magnetic north will also be updated, even though it is not being acted upon. When the car begins to move down the slope, into the depression, the angle of the camera is now pointing down, and the red ball is no longer in the visual field. Thus there is no red-directed disposition to override the magnetic pole-directed disposition, and the car begins to move in a different direction. This is a clear case of perspectival sensitivity in the full sense: maintaining and updating intentional dispositions as a result of changes in one's experiences, while those dispositions are being pre-empted by other dispositions of the system.

However, this is only an objection to perspectival sensitivity as sufficient for content if we find it unacceptable to attribute content to the more complex toy car. But not only does the complexity of the interacting dispositions of the modified toy car make the ascription more plausible; it also thereby severely restricts the range of physical systems which may be classed as perspectivally sensitive. And it is this last point which makes inroads against the panpsychist point, which was the only principled objection to the idea of the toy car having contentful states.

To complete this part of the project, then, one would need to show that an explanation of the Example has a use for content by showing that the Example involves the stronger notion of perspectival sensitivity with interacting dispositions. Unfortunately, this would require the results of empirical studies that have not, to my knowledge, been conducted. So far this chapter has made strong concessions, on the part of a priori, philosophical inquiry, towards empirical relevance. The point of this chapter is to motivate the belief that the notion of non-conceptual content developed in the following chapters is more than a mere formal possibility, in that it is also of explanatory use in empirical psychology. This motivation can be based on provisional findings, and does not require that the many empirical questions surrounding the Example be settled beyond a doubt.[8]

2.4.2 The "searching under a cloth" and "$A\overline{B}$" data

From the foregoing, it seems that one should be able to provide a contentful account of the Example. But now we get to the heart of the Example: it appears that a conventional (i.e. conceptual) content-based account cannot be given for the infant's behaviour. This conclusion proceeds in three stages: first I show how the two well-documented types of data, the "searching under a cloth" and "$A\overline{B}$" data, support the classification of infant behaviour as falling under the Example. Then I show how behaviour of the kind in the Example cannot be accounted for in objectual terms. This is enough to motivate the consideration of non-conceptual content. Later, in chapter 3, it is finally shown that conceptual content must be objectual, from which it may be concluded that a conceptual account of the Example is not acceptable.

A typical instance of evidence for the Example is the following.

Searching under a cloth: While playing with a toy, an infant will track the toy if it is moved in view of the infant; the infant will reach for the toy if it is in view. However,

> At about 4-5 months, the infant who has been playing with a toy and then seen an adult cover it with a cloth will do very little to recover the object. Even if a portion of the toy is left visible, sticking out from under the cloth, the baby will not lift the cloth. Two or three months later, the infant has progressed: if the object is fully covered, the infant will [still] do nothing, but if it remains partially visible, then the cloth will be removed, and the object recovered. Finally, at about 8-9 months, the baby spontaneously lifts the cloth even if the object is completely covered.[Harris, 1989, p 106]

Also, even after that stage, there is still evidence that meets the conditions of the Example:

$A\overline{B}$ data:

> As I noted earlier, the baby searched for an object that is completely hidden by a cloth at about 8-9 months. Even at this point, however, the baby continues to search inaccurately. For example, if the experimenter hides the object two or three times under the same cloth A and then on the next trial hides the object

[8] Campbell [Campbell, 1985] has criticized Peacocke's condition of perspectival sensitivity. But since Campbell's criticism focussed on the insufficiency of the condition for the possession of *concepts*, it supports, rather than undermines, what is being said here.

> under a different cloth B, the baby often makes what Piaget (1954) regarded as a very revealing [sic] error: The baby watches as the object is being hidden under the new cloth B but once it has disappeared, turns away, and searches at the cloth where the object was hidden on the initial trials, apparently expecting to find the object there. Occasionally, the baby will correctly approach the new cloth but if the object is not immediately discovered there will promptly revert back to the old cloth. This type of perseverative error has come to be called the $A\overline{B}$ error (i.e. A not B error). [Harris, 1989, p 112]

2.4.3 Two constraints on objectuality

When we think of something as an object, we think of it as object*ive*. That is, a) we think of it as a thing which exists independently of us, as a thing which continues to exist even while it is not being perceived by us. Also, b) we think about it as a thing which exists in only one place at a time, and we think of it as the kind of thing that has spatio-temporal continuity. (That is, we think of it as the kind of thing that, if it occupies place A at time t_1 and place B at time t_2, the object must trace out a continuous spatio-temporal path between t_1 and t_2 that connects A to B.)[9]

These need not be the only two constraints on objectual content. No doubt there are others, e.g. c) the subject must think of its perceptions of the object as the joint causal upshot of the object's position in the world and its own. But the two conditions given are enough for the purposes at hand: the "search under a cloth" data provide a case where an infant's contents concerning an object fail to meet condition a) for objectuality, and the $A\overline{B}$ data provide a case where an infant's contents concerning an object fail to meet condition b). This is now shown.

First, a terminological point should be made. Many psychologists, even (or especially) Piagetians, summarize the Example by saying that the child "lacks the object concept", or "lacks the concept of an object". Such phrasing would be misleading here, given that in this work I am sticking to a philosophical understanding of the term "concept". On that understanding, "possessing the concept of an object" would be the ability to predicate "object" of arbitrary particulars, and might have such co-requisites as possession of the concept *spatio-temporal continuity*, the concept *existence unperceived*, etc. These co-

[9] Thus, content that presents places as re-identifiable places is at least concomitant with, if not prior to, content which presents objects as objects. This concurs with [Campbell, 1994].

requisite concepts are quite sophisticated, and it is hardly contentious to claim that there is no evidence for such concepts in young children. I mean something much weaker than that when I speak of the ability to think of objects as objects, so in denying infants "the concept of object", I am making a much stronger claim. The lack of the ability to think of a toy *as* an object should not be construed as an inability to predicate "object" of the toy; rather, it should be seen as a limitation on the possible modes of presentation that the infant can use to think of the toy. Thinking of the toy *as* an object means thinking of the toy using a particular (type of) mode of presentation, one which has certain connections in inference, one that can be entertained even when the toy is not perceptually occurrent, etc.

To reiterate: in this context, when I say that we think of an object *as* a thing that is so-and-so, I do not mean that the concepts I use to spell out the so-and-so enter (as concepts) into the propositional attitudes of the thinker whose contents I am describing. When I say, for example, that a subject thinks of a toy as something which can exist unperceived, I am not attributing to that subject the ability to entertain contents about the toy of the form *The x is furry and the x is something which can exist unperceived.* The qualification "something which can exist unperceived" is not something external to the mode of presentation of the object, which is then combined with that mode of presentation to predicate something of the object. Rather, the qualification is internal to the mode of presentation itself; it indicates a type of mode of presentation.

More can be said to satisfy the sceptics. Specifically, more can be said about how these types are individuated. One proposal is this: for a mode of presentation to be of the type that presents something as that which can exist unperceived, it must be entertainable both when the object is, and when it is not, being perceived. As a first pass, this characterization is not bad, but it appears to be too weak. For surely if an infant can entertain a thought about a toy, say, then it can entertain that thought about the toy even when the toy is not present – the infant might be tricked by an illusion, for example. Thus, every mode of presentation of a perceptually occurrent object is trivially a member of the type; the distinctions between modes of presentation on which I wish to rely cannot be made if this objection holds.

There are two ways to respond to this observation: one is to deny that the infant's thought, if any, in the illusory case involves a mode of presentation of the toy; support from this would come from familiar externalist considerations. Also, a simplistic interpretation of the Fregean dictum that sense determines reference would seem to force a conclusion that the modes of presentation in the perceived toy/illusory cases must be different. If they are different, then a trivialization of the definition of the mode of presentation type under consideration has not been made. However, if one has a more relaxed (and more plausible) view of the Fregean dictum – that sense, together with the way the world is, determines reference – then one will not be allowed to infer that the modes of presentation in the two cases (toy/illusion) are distinct. The trivialization still threatens. In any case, even if the special case of the illusory toy could be addressed, there would remain the general possibility of trivialization by virtue of the infant being able to entertain a thought about the toy even when the toy is not present. Consider the possibility of a time lag between a toy's removal and the onset of an inability to employ the mode of presentation of that toy. We would not want to say that a mode of presentation of a toy, which normally can only be entertained when the toy is perceptually occurrent, is in the type under consideration ("presents the toy as that which can exist unperceived") just because of a time lag of that sort. Yet that is what we would be forced to do if we adopted the first pass at a characterization of that type of mode of presentation.

It is better, then, to make the second response: admit that the characterization of the type needs to be tightened. The mistake was to take the presence or absence of the toy reductionistically, to involve mere spatial proximity, being in front of the eyes, etc. Actually, we want to understand our ability to think of something as that which can exist unperceived as an ability to think of the object even though the object is not being presented in *perceptual experience* (even if it is spatially proximal). That is, the key is not the ability to think of the object when it is spatially absent, but rather when it is absent from experience. This is one place where the distinction between horizontal and vertical notions of non-conceptual content is useful (see chapter 3). The horizontal notion of non-conceptual content can provide an account of how perceptual experience presents the world as being, and this can be used in determining whether or not other, vertical contents are within the type under consideration. If the vertical contents do not fall within this class (i.e., if they are not entertainable, when the horizontal content of perception does not present the object), then such contents cannot be objectual contents.

2.4.4 Why objectual accounts fail

Now we can use these two constraints, a) and b), to see why different kinds of infant behaviour falling under the Example do not warrant an ascription of objectual content.

Searching under a cloth: There is no behaviour of the infant which suggests that it can have any thoughts at all about the object when it is not perceptually apparent. A fortiori there is no evidence that it can have thoughts, when the the object is not perceptually apparent, that involve the same mode of presentation as do those contents that the infant employs when the object is perceptually apparent. Therefore, the modes of presentation being used when an object is perceptually apparent fail to meet constraint a), and are thus not objectual.

$A\overline{B}$: The $A\overline{B}$ data seem to amount to a bald negation of objectuality constraint b). What could count as better evidence of failing to meet b) than that an infant looks at place A when they see an object disappear at B, with no plausible way for the object to have traced a continuous spatio-temporal path from B to A? Since the infant in this case fails to meet constraint b), the way it is thinking of the object cannot be an objectual way of thinking. Again, this is not to say that our objectual thoughts must meet some "internal" condition, must include the concept of being in one place at one time; rather, the claim is that a way of thinking of an object must meet some "external" condition, concerning, say, the conditions under which it may be entertained, in order for that way of thinking to meet constraint b). And the claim is that whatever those conditions might be, $A\overline{B}$ data seem a clear case of failing to meet them.

The conclusion that an objectual account is inadequate for accounting for the $A\overline{B}$ data is supported by Harris' analysis. After examining the failure of the various attempts that have been made to capture the way the infant is thinking in such cases, he finally suggests that these ways of thinking cannot be given a conceptual explanation:

> These very orderly data create a problem for any purely conceptual or cognitive interpretation. Consider, for example, the idea... that the infant

> does not appreciate that an object can be in only one place at a time, and fails to rule out A as a possible hiding place. Are we to say that the 8-month-old baby can rule out A for 3 seconds but no longer, the 9-month-old baby for 5 seconds but no longer, and so forth? If the baby has come to understand that an object can only be in one place, *why does it not apply this knowledge to delays of any length?* [Harris, 1989, p 115, emphasis added]

After suggesting a particular account of the data which does take non-systematicity seriously, he remarks:

> As soon as we talk in these terms, we are beginning to talk about the efficiency with which the baby processes the information, rather than simply proposing conceptual rules that the baby either understands or does not understand. We are admitting, at the very least, that a conceptual interpretation will not work.

Thus, Harris recognizes that conceptual accounts of the infant's mind won't do. Admittedly, it is unlikely that Harris is using the term "conceptual" in the same way that I, or philosophers in general, do. In fact, Harris seems to use "conceptual" to mean any intentional, content-involving account; a non-intentional, mechanistic account (he specifically proffers the proposals made in [Diamond, 1988]) is the best we can do. Although incorrect, his conclusion is understandable, since the option of *non-conceptual* intentional analysis is not in common cognizance. Without the awareness of this option, one might indeed equate "conceptual" with "contentful", and take the failures of the former to indicate the inapplicability of the latter.

Even if Harris doesn't actually mean to say that a conceptual account, philosophically understood, can't be given, it is in fact the case that the data do seem to preclude an *objectual* account. Harris' question, "why does [the infant] not apply this knowledge to delays of any length?" is the right one to ask.

The above arguments take a hard line in answering that question: demonstrating a need for non-conceptual content is achieved by showing that there is no correct objectual account of the Example. Perhaps this line is too strong; it seems to overstate the case to such an extent that it is susceptible to an objection. Specifically, it invites one to think of ways that an objectual account could be made to work for the Example. If such an account can be made plausible, the hard line is rendered impotent. For instance, in the case of the "searching under a cloth" data, some (including, it seems, Piaget) find it tempting to think in the following way. They admit that when the object is not being perceived by the infant, the infant cannot think of the object at all. But they then say that the infant, when the object *is* being perceived, is thinking of the object objectually, as something that can exist unperceived.[10] Presumably, these people would also wish to maintain that the infant that makes the $A\overline{B}$ error is thinking of the toy objectually, as a thing that must trace a spatio-temporally continuous path between A and B if it is at A and then later at B, but that the infant just forgets this fact, or forgets that the object was last seen at B, or some such. Furthermore, the data Harris discusses would have to be explained by saying that the 8-month-old baby can think objectually for 3 seconds but no longer, the 9-month-old baby for 5 seconds but no longer, and so forth.

This way understanding of content and objectuality is fundamentally flawed. It creates an intolerable gap between mind and world, between thought and embodiment, between content and the abilities which manifest that content. This Cartesian view of mentality has insurmountable problems, which, thanks to Wittgenstein, are now well-known (although obviously not well enough). Although the scepticism that this view inexorably yields constitutes a general reductio ad absurdum of the Cartesian position, it will be more persuasive to isolate the flaw in this particular application of that view. The attraction of the Cartesian view that our mental contents can outstrip their manifestation seems to derive from reasoning such as this: an adult could choose to act toward the object in a way identical to the infant in the "object under a cloth" and/or $A\overline{B}$ data, and yet still think of the object objectually. Surely (it is assumed) I have incorrigible knowledge of my own mental states; it seems that in such a situation that I am thinking objectually; therefore it must be so. If, despite my own display of the "searching under a cloth" or $A\overline{B}$ data, the ascription of objectual content is correct in my case, then so also it may be correct for the infant. So it has not been *shown* that the infant's contents *cannot* be objectual.

One obvious flaw in this is the assumption of one's authority concerning the objectuality of one's mental contents. But even if that self-knowledge is granted, the above reasoning is mistaken. It would perhaps work against a straw man version of a logical behaviourist position, where the identity of the local, macro behaviour of the (conceptually

[10] Compare [Evans, 1985, pp 259-260].

sophisticated) infant (and adult) is all that is taken into consideration when considering which mental states are ascribable. And indeed the behaviour of the infant and adult are identical, on that impoverished notion of behaviour. But on a less trivial account, there are differences of disposition and underlying mechanism, and macro differences of behaviour on a wider time scale, along with one's privileged access to the adult's mind but not the infant's, which together provide the warrant for ascribing objectuality to the adult, but not to the (similar with respect to local behaviour) infant. Together, these considerations permit one to maintain the hard line in the face of the Cartesian sceptic. The mistake of the latter is revealed to be thinking that the length of time that can pass for the infant to still search the correct place is somehow external to the way the infant is thinking of the object, when in fact, as one can see after cognitive significance is generalized to psychological significance, such facts are constitutive of our ways of thinking (see also chapter 3, section 3.8.2).

The advantage of the soft line is that a weaker claim is made, so the Cartesian objection doesn't ever arise, let alone require refuting. The soft line is to admit that perhaps an objectual story can be told for the infant in the Example, but it is inferior to a non-objectual story. The soft line proceeds through three stages. First, one must make sense of non-objectual content. Then one makes sense of an ordering of less objectual and more objectual ways of thinking. Finally, one applies some principle that justifies a preference of a more basic, less objectual account over a more complex, objectual one. Since even the first step of this process requires the results of chapter 3, this way of establishing the inappropriateness of an objectual account for the Example will be postponed until the end of that chapter. Thus, we have a situation concerning the failure of conceptual explanations of the Example that is symmetric to the one concerning the failure of non-contentful explanations of that same Example. Earlier, when establishing the contentful nature of the Example, it was pointed out that the *conclusive* way to show that the Example involves contentful phenomena is to provide a notion of content which can account for the Example. Yet if the general claim of this thesis is correct, this can only be done *after* the notion of non-conceptual content is fully introduced. Similarly, the *conclusive* way to show that a conceptual account of the Example is not acceptable is to show that a non-conceptual account for the Example can be given, in that it does not encounter the difficulties raised here, and that a simpler, non-conceptual account is to be preferred to a conceptual one. One further connection, then, will have to be shown in chapter 3: that conceptual content is objectual. Then the non-objectual content required for the Example may be understood in terms of the theory of non-conceptual content developed in that chapter.

2.4.5 Further empirical considerations

In my account of the Example, I adopted what many would call a Piagetian perspective. But it should be clear that one need not accept all aspects of the Piagetian story in order to use the Example to motivate a non-conceptual account. In particular, one need not be committed to the details that have received a large portion of the recent criticism: the particular age for which Piaget claimed certain abilities are and are not present, the claim that development proceeds in distinct stages, the claim that all infants follow the same sequence of stages, the claim that development is always in the direction of increasing objectivity, with no "backward" steps, etc.

Also, it should be mentioned that the interpretation of the data as falling under the Example, although part of the Piagetian orthodoxy, are now controversial, and are thought by many to have been discredited by recent research (e.g., [Spelke, 1985], [Diamond, 1988], [Baillargeon et al., 1985]). For example, Baillargeon et al. have collected data that have been interpreted as showing the following: that if even a two month old infant sees an object go behind a screen, the infant will exhibit surprise if the screen rotates back in a way that would be impossible if the out-of-view object were indeed behind the screen. This suggests that the infant has expectations about the possible ways the screen can move that take into account the position and shape of an unperceived object. That is, the infant appears to be thinking of the object as something which can exist unperceived, in my sense. Thus, there is no problem in ascribing an objectual content.

There are at least four responses to the doubts raised by such studies. First, the empirical controversy is indeed still a controversy, and several psychologists still question Baillargeon et al's interpretations, let alone their data. Second, I am not taking a stand on the exact age at which objectual accounts do become acceptable. Even if Baillargeon et al's data and interpretations are right for infants of two months, this does not directly support the applicability of objectual content in explaining infants younger than two months.[11]

A third rejoinder is to make a prior case for non-conceptual and/or non-objectual content from naturalistic considerations. Some might claim that although precisely *when* conceptuality/objectuality sets in is an empirical fact, it is a naturalistic requirement that *if* it sets in, it can only do so after a period of non-conceptuality/non-objectuality. A variant of this rejoinder is given by Cussins [Cussins, 1992]. One way of applying his point is to say that even if *all* infants think conceptually, there *must*, given a naturalistic continuity for which natural selection is applicable, be some stage of the development of that infant that is pre-conceptual, whether that stage be in the womb, or even in the phylogenetic history of our species. The idea is that if there once were no concepts, and now there are, there must have been an intermediate non-conceptual yet contentful state.

It should be clear that I do not think such a rejoinder will work, for if I did, I would no doubt rely on it to motivate an investigation into non-conceptual content rather than looking at the Example. In particular, I think the rejoinder is too powerful. Perhaps it is right to make a philosophical case for a *distinction* between content-warranting abilities and concept-warranting ones. But it would be to overstate the case to rule out a priori the very real empirical possibility that content-warranting abilities did (or even generally do) happen to appear at the same time as concept-warranting ones. And it is the latter possibility that would undermine the *empirical* need for a notion of non-conceptual content.

The fourth possible reply to the Baillargeon et al. data is both more modest and more compelling. That an infant's behaviour meets the objectuality constraints in one situation does not warrant ascription of objectual content in all other cases. Even if the infant, in the context of the rotating screen, behaves in a way consistent with objectuality, it could still be the case, for the various reasons given, that the infant lacks objectuality in other contexts, even those involving the same object. It is important to make clear that

[11] Admittedly, if I were to concentrate on these younger infants as a fall back position, I would have to take more care concerning the tracking data used to motivate a contentful account of the Example in the first place. If it just so happened that this tracking ability only developed at two months or later, then it would render this second response to Baillargeon et al. impotent.

a non-conceptual account is not refuted if conceptual constraints are sometimes met; on the contrary, it requires a non-conceptual account to explain cognition that sometimes involves objectual, and sometimes non-objectual modes of thought concerning the same item.

CHAPTER 3

Structure, Concepts & Objectivity

3.1 Introduction

As stated before, it is my contention that content-based explanation of the Recalcitrant Phenomena may be achieved with the assistance of a notion of non-conceptual content. To make my notion of non-conceptual content more precise, I now turn to a discussion of the nature of content constituents, and concepts. I supplement an account of what it is for something to be a constituent of a content with conditions that must be met for a constituent to be a concept, and then argue that not everything which meets the generic conditions for constituency also meets the conditions for conceptuality. Thus, there is content with constituents that are not concepts: non-conceptual content.

3.2 Structure

In saying that a content has structure, I mean that it has constituents. A content is composed of its constituents. The constituent of a content is itself a content, though typically a partial one.[1] Content constituents may also be *possessed* by thinkers. For example, a concept is (a type of) content constituent, yet we also speak of possessing the concept C. This just means that one possesses an ability which, when combined with certain other abilities, confers the ability to entertain contents that have C as a constituent. More precisely, I stipulate as part of what is meant by a constituent of a content, the Possession Principle:

[1] Recall my terminology from chapter 2: whole contents have truth values; partial contents must be combined with other contents to yield a content with a truth value.

> **Possession Principle [sufficiency]:** Let C be a set of content constituents, and let $X(C)$ be a content which has the members of C as constituents. If a thinker possesses each of the constituents in C, and grasps the significance of their mode of combination, then the thinker possesses the ability to entertain the content $X(C)$.

If the above is viewed as providing sufficiency conditions for the ability to entertain a content $X(C)$, there is a complementary principle concerning necessary conditions:

> **Possession Principle [necessity]:** Let C be a set of content constituents, and let $X(C)$ be a content which has the members of C as constituents. If a thinker possesses the ability to entertain the content $X(C)$, then the thinker must also possess each of the constituents in C (and grasp the significance of their mode of combination).

This second Principle is employed and discussed in section 3.3.3 and in chapter 5, section 5.3.1.

The structure of a content has implications for every other aspect of that content, including its psychological significance (e.g., the validity of an inference depends on the structure of the contents involved), its semantic value (the semantic value of a content depends on the semantic value of its constituents), and the conditions for a subject to be able to entertain it (grasping a content with constituents confers an ability to grasp related contents: contents composed of those constituents).

It is a common belief that it is part of the definition of content and concepts that the former is composed of the latter. That is, any constituent of a content is a concept. Witness Crane:

> Now many think it obvious that the contents processed by the visual system do have constituents. For the theory assigns structure to these states, analysing them – as it may be – in terms of concepts such as that of a *zero-crossing* (Marr 1982). *But if a concept just is a constituent of a content*, then the constituents of these computational contents will be concepts *by definition*... So the mere idea of a content that is not composed of concepts does not help to explain the idea of non-conceptual content [Crane, 1992, p 140-141, emphasis added].

The degree to which this mistake of defining concepts as the constituents of content is tempting is demonstrated by the fact that Crane makes it in the very context of writing about *non*-conceptual content. It is because Crane accepts this weak notion of what it is to be a concept that he rejects the notion of non-conceptual content that I advocate (content that is not composed of concepts) in favour of a different notion: content the possession of which does not require the subject to possess the concepts used to specify it (see section 3.3.1).

A more substantial notion of a concept is required, yet one that is still in line with the uses to which concepts are typically put (e.g., the Generality of concepts explaining the productivity of thought) and the claims that are typically made of conceptual content (e.g., that grasp of such contents requires possession of the concepts that appear in the content's canonical specification). I provide a substantial account of what a concept is, beginning in section 3.3. It is a consequence of this more substantial account of what concepts are that Crane's objection to the simple notion of non-conceptual content is rendered ineffective. Since being a content is more than just being a content constituent, room is made for content constituents that are not concepts.

Must content have structure? If not, then the existence of non-conceptual content would be established directly: a content with no constituents is one, a fortiori, without conceptual constituents. Of course, someone like Crane, who takes concepts just to be constituents of content, will think that non-conceptual content must be content without structure:

> So perhaps the idea is this: just as conceptual content is content that is composed of concepts, so non-conceptual content is content that is not composed of concepts. Conceptual content is 'structured' content, and non-conceptual content is 'unstructured' content. [Crane, 1992, p 140]

One need not object to the possibility of *some* contents being unstructured. The conceptual constituents of a content are themselves contents, and yet they have no further structure (at least not in any conventional sense of "structure"; but see next footnote). However, the notion of a whole (as opposed to partial) content without any structure, which is what Crane is considering, should be resisted. It should be resisted, and not only because we would have to make sense of a) a whole content that has no inferential connections to other contents (except for contents that have the unstructured content as a constituent), b) a content the grasping of which has no implications for the grasping of other contents, and c) a content which is semantically unrelated to any other content.[2] More importantly, it seems the very possibility of a content being that which can be true or false requires that the content have some constituent structure. It is in virtue of the semantic values of these constituents matching or not matching that the content is true or false. In the most fundamental case, to be false is to predicate something of one or more particulars, which is not true of those particulars. To predicate something false requires two tasks to be achieved: the introduction of a particular, and a predication of a property to that particular. Inasmuch as a content achieves both of these, it has structure.[3]

The structure of content is explanatorily prior to its other properties. This can be maintained even if one believes that the structure of content is epistemically posterior to those other properties. According to that view, when attempting to explain the behaviour of a subject, we are not first presented with the contents of the subject, much less with the structure of those contents. Rather, we (at least some of the time) are confronted with the non-contentful properties of the subject, including behaviour, and we ascribe the contents that best explain those behaviours. Whether we ascribe a content with this structure or that will depend on which is the best explanation of the data we have so far, and to which explanations/predictions we wish to be committed in the future. But once ascribed, the content's structure explains the very data that we used to ascribe it. The ascription of structure to content is an instance of inference to the best explanation.[4]

[2] One might think that contents can be unstructured and yet have, e.g., inferential connections to other contents. For example, *bachelor*, although unstructured, has inferential (indeed, logical) connections to *unmarried male*. It is a consequence of the position I am adopting here that inasmuch as such connections do exist between contents, those contents thereby are not unstructured, but contain some implicit structure. This point holds for other contents that one might ordinarily take to be unstructured, such as Peacocke's scenario contents [Peacocke, 1993, ch 3]: inasmuch as they have inferential connections, they have an implicit structure. But note that this position is not crucial to my account of non-conceptual content; it is adopted only to motivate the idea that non-conceptual content has structure.

[3] The foregoing assumes the case of whole contents with simple subject-predicate structure. And of course, not all whole contents have this structure (e.g., quantified contents). But this does not affect the point being made, since these contents must likewise have some structure in order to have a truth value. And in fact, their structure is explained in terms of basic subject-predicate structure.

[4] Of course, some (including simulation theorists in psychology [Harris, 1991], Quine [Quine, 1960, p 219] and even, as Peacocke points out, Vico [Peacocke, 1993, p 169]) would not be happy with the account

To illustrate how structure can be inferred as the best explanation of other properties of content, I offer the following. This illustration assumes that content is ascribed by means of a two-stage process: first the ascription of contents with certain inter-relationships (such as content a implying content b, or knowing what it is for a to be true implying knowing what it is for b to be true), without any postulation of the structure of those contents; second, an ascription of structure to those contents which in turn explains the inter-relationships postulated in step one. The illustration assumes step one has already been completed, and details a case of the ascription of structure in step two. Admittedly, this two-stage process is rather artificial. In actual practice, the ascription of structure could very well be interleaved with the ascription of the other properties of content. But this does not require an alteration to the conception of structure I am elucidating.

Consider the eight distinct contents A-H. These contents are not independent of each other; the dependencies and their direction can be notated with $\Rightarrow$. Although an example could be constructed for any of the essential properties of content, assume, for sake of concreteness, that A $\Rightarrow$ E means that A implies E. Suppose the following fourteen dependencies exist between the contents:

A $\Rightarrow$ E
A $\Rightarrow$ D
A $\Rightarrow$ G
A $\Rightarrow$ H
B $\Rightarrow$ C
B $\Rightarrow$ F
B $\Rightarrow$ G
B $\Rightarrow$ H
C $\Rightarrow$ G
D $\Rightarrow$ H
E $\Rightarrow$ G
F $\Rightarrow$ H
C, F $\Rightarrow$ B
D, E $\Rightarrow$ A

I present here of how we come to know other minds. For them, knowledge of the structure of another's content is not inferred from the joint consideration of that other's behaviour and a psychological theory. This is no objection to the point I am making, however, since I am arguing for the explanatory priority of structure in the face of epistemic posteriority. If structure is also epistemically prior to (or concomitant with) its other properties, as the simulation theorist would argue, this would only serve to further justify the emphasis I am placing on structure in my account of conceptual and non-conceptual content.

On analysing this set of relationships, it becomes clear that there is some hidden structure. Indeed, the fourteen statements of dependency jointly give (partial) expression to the each of the eight contents, even if it is not in the standard way of using 'that' clauses. The prior (implicit) structure allows us to speak of a particular content or set of contents; only after the contents are so individuated may we then make explicit this prior structure. This is done, as said before, by inference to the best explanation. A more economical way of capturing these relationships, a way that explains them, is to replace the fourteen dependencies with only eight equations:

A = P(a) & Q(b)
B = P(b) & Q(a)
C = P(b)
D = Q(b)
E = P(a)
F = Q(a)
G = Ex P(x)
H = Ex Q(x)

where "=" means "has the structure of".

The fourteen relations above can then be explained as consequences of these eight postulates, along with the standard interpretations of conjunction and the existential quantifier. Of course, the fourteen relationships are only a small fraction of all the logical dependencies that are constitutive of the contents in question; the ascription of structure to these contents becomes irresistible as one attempts to accommodate more and more of these relations.

The points just made concerning structure and implication relations between contents also hold for structure and other constitutive relations between contents. For example, possession or entertainment relations between contents (e.g., if one has the the ability to entertain contents A and B then one thereby has the ability to entertain content D) also generate an implicit structure, as do semantic relations between contents.

3.3 Concepts

With a clear notion of what it is for something to be a constituent of a content, it can be determined what it is for a constituent to be a concept. Then showing that the conditions for being a constituent do not imply the conditions for conceptuality will allow for the possibility of constituents which are non-conceptual.

As pointed out in section 3.2, it would be a mistake to *define* concepts as being just content constituents. Rather, there seem to be some further constraints that are essential to concepts, which do not follow directly from the mere fact that they are content constituents.

3.3.1 Horizontal vs. Vertical content

One point of universal agreement concerning conceptual content is that it has the following property: to grasp a conceptual content, one must possess the concepts employed in the canonical specification of that content; to grasp a non-conceptual content, one need not possess the concepts employed in the canonical specification of that content; e.g., [Peacocke, 1986, p 17], [Crane, 1992, p 142], and [Cussins, 1990, p 382 ff].

Highlighting these properties of content is useful[5] for the purpose of investigating a kind of non-conceptual content which is distinct from the kind developed in this thesis. That is, it is useful when considering what I call *horizontal* non-conceptual content, which has its primary application in the theory of perception (see the discussion in chapter 4, section 4.1). The primary difference is that unlike the *vertical* notion of non-conceptual content which I am considering, horizontal non-conceptual contents are not possible objects of the attitudes. Rather, their purpose is to allow the individuation of concepts to proceed by appealing to how one's dispositions to judge are connected with what is given in experience. The term "horizontal" is meant to evoke a connection with conceptual content at a given point in time, moving from the outer, experiential realm toward the inner realm of the attitudes. Conversely, "vertical" non-conceptual content makes (some of its) connections with conceptual content over time, e.g. in development, or concept acquisition, and, like conceptual content, is internal to the attitudes.

The statement of the properties speaks of "entertaining" a content, rather than taking an attitude toward it, because horizontal content is meant to capture the content of perceptual experience that is prior to any doxastic states. The case of persistent illusions,

[5] So useful, in fact, that the properties are often taken to be *definitional* of the two types of content.

such as the Müller-Lyer illusion [Müller-Lyer, 1981], is often cited in order to support this two-factor approach to content. There is, on the one hand, the content of perceptual experience, which is "given" prior to judgement, is (at least partly) non-conceptual, and which is "entertained" by a person, whether they are aware of the actual lengths of the lines or not. This is contrasted with the content of one's thoughts and judgements concerning the lines, which are presumed to be conceptual, and which typically will depend on what one knows about the illusory nature of the lines involved.

As stated before, despite the utility of acknowledging these properties of content when considering horizontal non-conceptual content, I will not define conceptual and non-conceptual content in terms of them. Rather, I explain (in section 3.8.1) why more conceptual and non-conceptual content have these respective properties, by appealing to more fundamental, defining aspects of content.

3.3.2 The Generality Constraint

Another well-known characteristic of conceptual content is captured by what Evans has called the Generality Constraint [Evans, 1982, p 104]. Peacocke states it thus:

> **Generality Constraint** If a thinker can entertain the thought Fa and also possesses the singular mode of presentation b, which refers to something in the range of objects of which the concept F is true or false, then the thinker has the conceptual capacity for propositional attitudes containing the content Fb. [Peacocke, 1993, p 42]

I do not question the Generality Constraint: it is necessarily true of all concepts. In fact, it is the very fact that concepts must meet the Generality Constraint that makes them inappropriate constituents for the contents needed in order to provide psychological explanations of certain phenomena. The Generality Constraint applies to any content Fa that is composed of the concepts F and a. However, I argue that there are contents that are not composed of any concepts F and a (and not because they have some other conceptual form, such as being quantificational, etc.). With respect to these contents, the Generality Constraint is silent. Therefore, one can uphold the Generality Constraint and yet assert that there is content whose constituents do not universally recombine.

Despite the fact that there is near-universal agreement on the truth of the Generality Constraint, it would be a mistake to *define* concepts as content constituents for which the

Generality Constraint holds. There are three reasons why one should not.

First, it might only be necessary, and not sufficient, for a constituent to recombine arbitrarily with other constituents within the appropriate range, for it to be a concept.

Second, the Generality Constraint presupposes a well-defined notion of concept; it certainly was not intended to define what concepts are. Indeed, there is reason to believe that one cannot even *state* the Generality Constraint without employing a prior notion of concept (consider: what are the entities that are supposed to recombine? and what is a conceptual capacity?). If this is so, the Constraint could hardly be used to distinguish concepts from non-conceptual constituents of content.

Last, it seems that the Generality Constraint is derivable from other, defining characteristics of concepts. If one can identify these other properties of concepts, from which one can *derive* the Generality Constraint, then one will have a much better idea of what the essential properties of conceptual constituents of content are.

3.3.3 Peacocke's derivation of the Generality Constraint

This is precisely what Peacocke has done, with his *referential* explanation of the Generality Constraint [Peacocke, 1993, pp 43-44].[6] The idea here is to use Peacocke's derivation to make clear just what it is for a constituent of a content to be conceptual (what it is about a constituent that guarantees, inter alia, that the Generality Constraint holds of it). This will, in turn, make it clear just what a non-conceptual constituent of content must be, by making clear what it is not.

Some might be tempted, in the retreat from stipulating that the Generality Constraint is true of concepts, to seek an empirical justification of it. The most well-known attempt at this is Fodor and Pylyshyn's [Fodor and Pylyshyn, 1988]. An empirical justification may take one of two forms. First, one can assume that all content is conceptual, and

[6] That it is a referential explanation should not be taken to imply that it explains why conceptual contents meet the Generality Constraint (and why non-conceptual contents do not) in terms of a difference in reference between the two kinds of content. This would prohibit the unitary notion of reference on which I insist (see section 3.3.4). Rather, the differences between the two kinds of content are generated by a requirement, only in the case of conceptual contents, for the subject to *know what it is for* something to be the reference of a concept.

then attempt to show empirically that *all* content meets the Generality Constraint. The problems with that attempt are pointed out in chapter 7, section 7.1. The other form an empirical justification may take is to admit that not all content is conceptual, and provide some criterion by which one might distinguish the conceptual and non-conceptual. Then one attempts to show only that the conceptual contents meet the Generality Constraint. This less ambitious justification would avoid the problems detailed in section 7.1, but in assuming a criterion by which one can distinguish conceptual from non-conceptual content, it is powerless to assist in the current task of providing such a criterion.

In section 3.2, I noted the Possession Principle: possession of a content constituent C, when combined with the possession of the co-requisite constituents, enables one to entertain all contents that have C as a constituent. Although it might appear otherwise, the Principle does not imply the Generality Constraint; the latter is stronger. Provided that a thinker can entertain the thought Fa and possesses the concept b, the Constraint guarantees that, as long as b refers to something about which F can be true or false, the thinker can entertain Fb. A fortiori, the Constraint implies that given such circumstances, there is a content Fb to entertain. But there is no such implication of the Possession Principle. It only states that the thinker can think Fb *if Fb is a content*; it does not, like the Constraint, guarantee that there is a content Fb. How could it be that the predicational combination of F with b could fail to be a content? We already have one answer to this question to hand: Fb might be *semantically anomalous* [Cussins, 1990, p 417]: it might be that b does not refer to something of which F can be true or false. But the Constraint explicitly rules out such cases, so if the Constraint is stronger than the Principle, as I am claiming, there must be another way that Fb could fail to be a content. I return to the task of making this clear in section 3.4.

Thus, we arrive at a conclusion Peacocke has reached earlier: the Generality Constraint should not be stipulated as definitional of concepts, nor should it be justified empirically. How then should it be explained? Peacocke provides a derivation that is neither stipulative nor empirical; it will be useful for the purposes at hand to make explicit all the steps in his derivation:

> 1) The subject is able to entertain the thought Fa, and possesses the singular sense b (antecedent of the Generality Constraint).

> 2) "Since the thinker is capable of entertaining the thought Fa, he possesses the concept F" ([Peacocke, 1993, p 43]; definition of content constituents/Possession Principle).

Peacocke does not take this premise to hold without restriction. Earlier, he denies that "a thinker who has attitudes to thoughts containing the concept *red* must fulfill its possession condition" [Peacocke, 1993, p 28]. This is because he recognizes, for Burgean reasons [Burge, 1979], the possibility of the partial grasp of a concept. Thus, at the beginning of the chapter which contains this derivation of the Generality Constraint, he explicitly restricts himself to the special case of fully grasped concepts [Peacocke, 1993, p 41]. From this, one might conclude, as Ludwig does, that Peacocke intends that the Constraint applies "only to thought of thinkers when they fully grasp the concepts in those thoughts" [Ludwig, 1994, p 481]. However, in Peacocke's reply to Ludwig, he claims that "the systematicity for the Burgean attributions follows from systematicity for the cases of full understanding, on the account I gave" [Peacocke, 1994b, p 497].[7] Unlike Peacocke, I will take premise 2) to hold without restriction. In the only case where it matters (chapter 5, section 5.3.1) I explain how I handle the Burgean considerations. Returning to Peacocke's derivation:

> 3) Therefore, the subject possesses the concept F (by 1 and 2).
>
> 4) Possessing a concept is knowing what it is for something to be its semantic value (the Identification [Peacocke, 1993, p 23]; argued for by Dummett [Dummett, 1981, ch 3]).
>
> 5) Therefore, the subject knows what it is for something to be the semantic value of F (from 3 and 4).
>
> 6) If a subject knows what it is for something to be the semantic value of a concept F, then the subject knows what it is for an arbitrary object to fall under F (implicit).

Peacocke does not provide an explicit motivation for 6). But some idea of why he might take it to be true can be gleaned from his consideration of the special case of taking the semantic value of a concept F to be a property:

[7] Thus, it must be that Peacocke would now retract what he said earlier: "the reason for these restrictions is that the phenomena discussed here hold without qualification only for such attributions" [Peacocke, 1993, p 41].

> If the semantic value of a concept is a property, then under the Identification, grasp of a concept, meeting its possession condition, is knowledge of what it is for a property to be the semantic value of a concept. This may make it sound as if such knowledge falls short of knowledge of what it is for something to fall under that concept. But it does not. Included in the knowledge of what it is for a property to be the semantic value of a concept F is knowledge that for an object to be F is for it to exemplify the property that is its semantic value. This is just an expansion of what is involved in a property being the semantic value of a concept.

Although what Peacocke says here is not intended to be an argument for 6) (and despite the limitation of the passage to the special case of properties as the semantic values of concepts), we can consider what would have to be true in order for 6) to be supported by what he says here about "knowing what it is for". It only does so if "knowledge that for an object to be F is for it to exemplify the property that is its semantic value" implies "knowledge of what it is for an arbitrary object to fall under F". One might argue that this is not the case along the following lines:

> Even assuming that "knowledge that for an object to be F is for it to exemplify the property that is its semantic value" is meant to be knowledge concerning an arbitrary object, it seems I can have that knowledge about any concept, even one I do not possess. Suppose "glorkish" expresses a concept, *glorkish*. I know that for an (arbitrary) object to fall under the concept *glorkish* is for it to exemplify the property that is the semantic value of *glorkish*. But that does not imply that I know what it is for an arbitrary object to fall under *glorkish* in any sense that will guarantee my ability to think thoughts involving *glorkish*.

Seeing what is wrong with this line of reasoning tells us a little more about concepts. It is true that I know that for an (arbitrary) object to fall under the concept *glorkish* is for it to exemplify the property that is the semantic value of *glorkish*. But this is not the kind of knowledge Peacocke has in mind in the passage cited. The knowledge I have concerning *glorkish* is theoretical, philosophical, conceptual knowledge about philosophical logic. By "knowledge that for an object to be F is for it to exemplify the property that is its semantic value", Peacocke means something like "knowledge that for an object to be F is for it to exemplify P" where P is some non-theoretical, fundamental way of thinking of the property which is the semantic value of F. For example, for the case of the concept *red*, what is required is not "knowledge that for an object to be *red* is for it to exemplify the property that is the semantic value of *red*", but rather "knowledge that for an object to be *red* is for it to exemplify redness".[8] And I do not possess this knowledge for the case of *glorkish*, nor any other concepts I do not possess.[9]

Despite the failure of this direct argument against 6), it at least calls to our attention the fact that finding support for 6) is important, for it is precisely in 6) that the move is made from the ability to think one thought to the ability to think an entire range of thoughts. Yet there does not seem to be a clear explanation for why 6) is true. Before considering this further, I will continue the presentation of Peacocke's derivation:

> 7) The subject knows what it is for an arbitrary object to fall under F (5 , 6)
>
> 8) The subject knows what it is for something to be the semantic value of b (1 and 4).
>
> 9) If a subject knows what it is for something to be the semantic value of a singular sense b, it knows what it is for an arbitrary object to be the referent of b (explication of "knows what it is for"; see [Peacocke, 1993, p. 22]).[10]

[8] Compare Evans [Evans, 1982, p 110]: "Evidently a subject cannot be credited with an Idea a unless he knows what it is for a proposition of the form $\ulcorner \delta = a \urcorner$ to be true." Evans uses the variable δ to indicate a *fundamental* Idea of an object: a way of thinking of an object which presents it as possessing that which differentiates that object from all others. This is similar to the way of thinking that Strawson requires for there to be a unified framework within which to locate particulars and thus be capable of conceptuality; see section 3.6).

[9] There is another question: why doesn't the Identification 4) imply for properties the same thing that it implies for particulars? That is, the Identification implies that a subject that possesses the singular sense b knows what it is for an arbitrary object to be the referent of b (see step 9)), below). Thus, why isn't it the case that "knowing what it is for something to be the semantic value of F" (the antecedent of 6)) implies not the consequent of 6), but instead: that the subject knows what it is for an arbitrary property to be the semantic value of F? I think the only answer to this is that the asymmetry is part of the constitutive difference between subject and predicate. That is, it is constitutive of a content constituent F being a predicate concept that knowing what it is for something to be the semantic value of F requires that the subject know what it is for an arbitrary object to fall under F. Accordingly, it is constitutive of a content constituent a being a subject concept that knowing what it is for something to be the semantic value of a requires that the subject know what it is for an arbitrary object to be the referent of a.

[10] Why is it not the case that a subject could know what it is for a particular thing, B, to be the referent (semantic value) of b, and yet not know what it would be for an arbitrary object to be the referent of b? This can only be answered later, in section 3.5.

10) The subject grasps the semantic significance of the mode of combination of F and b (stipulated of concepts).[11]

11) If, a) for each constituent of a thought, a thinker knows what it is for something to be its semantic value; and b) the thinker grasps the semantic significance of the mode of combination of the constituents of the thought, then the thinker is in a position to know what it is for the thought to be true (explication of what it is to grasp the semantic significance of the mode of combination of the constituents of a thought).

Now either 11) is too strong, or else some other parts of the derivation are unnecessary, namely, 6), 7) and 9). For it seems that 5), 8), 10) and 11) will imply what was to be demonstrated, namely:

12) the thinker is in a position to know what it is for Fb to be true.

Yet 5), 8), 10) and 11) do not depend on 6), 7), or 9). Also, 11) is hard to motivate, for reasons similar to those mentioned in the discussion of 6) and 9). A replacement for 11) that actually uses 7) and 9) would be:

11′) If, a) a thinker knows what it is for an arbitrary object to fall under F, and knows what it is for an arbitrary object to be the referent of b; and b) the thinker grasps the semantic significance of the mode of combination of the F and b, then the thinker is in a position to know what it is for the thought Fb to be true (assumed).

This seems like more of a referential explanation of the Generality Constraint, rather than a mere restatement of it, and is very much in the referential spirit of Peacocke's derivation as a whole. Furthermore, the assumptions it requires one to make are weaker and therefore less contentious.

3.3.4 Concepts and reference

The results of Peacocke's derivation of the Generality Constraint can now be applied. There are two relevant possibilities:

[11] See p 24: "...the role of a first-level concept in predicational combination is essential to it and is given automatically in its individuating possession condition. This is only to be expected if, as is plausible, there is no such thing as possessing a concept while failing to grasp its significance in predicational combination." This is repeated on p 44.

1) On the traditional notion of content, sense determines reference; content determines a semantic value. It is because concepts are stipulated to be those constituents (of whole contents) which have objects and properties as their semantic values that Peacocke can show that they are recombinable, that the Generality Constraint is true. For it is essential to the very idea of a property (the semantic value of a predicative concept) that it can hold of any particular within its range of application; and it is part of the very idea of a particular (the semantic value of a singular concept) that any property, in which range of application the particular falls, can hold of it.

2) The defining characteristic of concepts that implies the Generality Constraint is the Identification: the claim that to possess a concept is to know what it is for something to be its semantic value. It is because concepts are stipulated to be things for which the Identification must hold that the recombinability of concepts is ensured. For, even considering a constituent that has an objective property (or particular) as its semantic value, if it were not the case that grasping a thought containing that constituent entailed knowing what it is for something to be the semantic value of the constituent, then the thinker would not be guaranteed the knowledge of what it is for an arbitrary object to fall under that property (or of what it is for an arbitrary object to be the referent of the singular sense). Thus, possessing F and b would not always entail the ability to think Fb.

Since the intention is to use an account of concepts to help clarify the notion of a non-conceptual constituent, the choice of 1) or 2) (or both) appears crucial. It is my contention that 2) is to be preferred to 1) as a stipulation of what is characteristic of concepts. This may be surprising; my interests in using the notion of non-conceptual content to explain pre-objective cognition might seem to be better served by opting for 1), with its suggestion that non-conceptual contents have non-objective semantic values. Thus, an explanation for my choice of focussing on 2) is in order.

I do agree that the requirement in 1) holds for concepts. That is, the semantic values of concepts must be objective properties and particulars. But that doesn't help define concepts in particular, since I think the requirement in 1) holds for *all* (partial) constituents of content in general. Although I do not think that intelligibility requires that there only be one conceptual scheme [Davidson, 1974], I do think that it requires that there be one realm of reference, one world which we all share. Yet consider my (presumably objective) contents about a glass, and an infant's (arguably pre-objective; see chapter 2) contents that we would normally say are about the same glass. Why should we recognize two (at least!; consider the fact that anyone else might think about the glass, and the infant might have contents about the glass on more than one occasion, which would presumably require, on the view I am criticising, the postulation of a new entity each time) entities in the realm of reference, when only one is required?

Of course, there are many cases in which thoughts that are, in some superficial sense, "about" the same thing, are, more strictly speaking, about different things. For a linguistic parallel, consider: "that statue" and "that lump of clay". In casual discourse, we could say that the two terms, in appropriate conditions, refer to the same thing. But in fact, the two terms have different referents: the referent of the first has persistence conditions which are distinct from those of the referent of the latter term.[12] Similarly, I do not wish to claim that in all cases where conceptual and non-conceptual contents superficially appear to be about the same thing, they actually are. Rather, the notion of a unitary realm of reference is this: if something is the referent of a non-conceptual content, it is also available to be the referent of a conceptual content.

I think many people are tempted to embrace the notion of what we might call "subjective realms of reference" because of a lack of confidence in, or ignorance of, the idea that the realm of sense can account for all the differences between, on the one hand, how adult humans experience the world, and on the other, how infants (or, in other contexts, animals, people of other cultures, etc.) experience the world. Yet these differences *can* be accommodated by a difference in content, a difference in how the *same* world is presented to these subjects.

Not only are subjective realms of reference not needed, they make it more difficult to explain the very developmental/temporal aspects of intentionality which are my primary reason for investigating non-conceptual content in the first place. Consider the development of a scientific theory: just as an understanding of scientific progress is hindered by taking "electron" to have been referring to different things at different stages of the de-

[12] Thanks to Chris Peacocke for bringing this point to my attention.

velopment of physical theory [Putnam, 1975, pp 235-238], so also will it be a hindrance to think of the pre-objective thoughts of an infant as referring to some pre-objective "smear" when trying to understand the development of the infant's ability to think about the world. On the view I am advocating, then, non-conceptual content does not give complete access to *a* less-than-objective world, but it instead gives less-than-complete access to *the* objective world.

Thus, 1) is not an attractive option; defining concepts as those constituents with objects and properties as their semantic values would seem to rule out the possibility of non-conceptual content. 2) must be taken as giving the defining characteristics of concepts.

However, 2) relies heavily upon the problematic notion of "knowing what it is for" something to be the semantic value of a constituent. Before more can be said about this distinct form of knowledge, the notion of a non-conceptual constituent must finally be explicated.

3.4 Proto-concepts

The strategy all along has been the following: to make clear what it is for a constituent of a content to be conceptual, and then use this account to explain what it would be for a content to be composed of constituents that are not concepts. I take it that part of an account of what makes a content conceptual is an account of why it is General. Admittedly, one can conclude that a content is non-conceptual from the fact that it is non-General, even if one does not know why conceptuality implies Generality. But the further task of explaining is vital here. If I were simply to use the premise "if a content is non-General, then it must be non-conceptual" without explication, there might be worries that the premise is only vacuously true, because the antecedent is impossible. The explication I give now makes it more understandable how the premise can be non-vacuously true (i.e., by explaining how a content could fail to be General).

In section 3.3.3, we saw that Peacocke's derivation of the Generality Constraint depended upon taking concepts to be those contents the entertaining of which requires a subject to know what it is for something to be the semantic value of the content. Call a content with this property a *C-content*. Rather than stipulating concepts to be C-contents, it would be better to understand what concepts are in terms of the essential properties of contents – cognitive significance and semantic value – directly. That is, it is better to define concepts as those contents whose cognitive significance and semantic value imply that they are C-contents. What we need now is an explanation of how the fact that a content is a C-content can be a consequence of a content's cognitive significance and/or semantic value.

If there is to be a possibility of non-conceptual content along the lines I am pursuing, then whether or not a content is a C-content cannot depend *solely* on the semantic value of the content. This is because, as I have argued, non-conceptual contents must have the same semantic values as conceptual contents. Non-conceptual contents provide a different access to the world, not access to a different world. The other possibility for distinguishing C-contents from other contents, and therefore the conceptual from the non-conceptual is distinguishing on the basis of a difference in cognitive significance. It is shown in the section 3.4 that a consequence of this is that to make sense of non-conceptual content, one must make sense of a *gap* between the semantic value and the cognitive significance of a content.

Before moving on, we can ask: must there be both non-conceptual subjects and non-conceptual predicates? Perhaps a failure of Generality could be explained with respect to only one of these two kinds of constituent. Perhaps, e.g., there are only conceptual subject contents, with objective particulars as their semantic values. If so, all non-conceptual contents would be either partial contents that are predicates, or whole contents that are the composition of a non-conceptual predicate content and a conceptual subject content.

Although a technical possibility, an account of non-conceptual content that does not permit the notion of a non-conceptual subject content is surely too impoverished for the explanatory purposes at hand. For there is the very real empirical possibility that the patterns in the development of a non-conceptual phenomenon will demand the postulation of a non-conceptual way of thinking of an object. Recall one of the intended applications for non-conceptual content: the intentional explanation of an infant's object-directed behaviour prior to the development of object-permanence. Independently of the actual facts about the development of object permanence, consider the following possibility. Suppose an infant can think thoughts involving a range of predicates (e.g., ones we might specify or attempt to specify with the concepts *high, low, to-the-right, to-the-left, etc.*), of objects in view, be it the infant's mother, or a toy, or a cat. Now suppose that prior to a particular learning experience, the infant is unable to apply any of the predicates to an out of view toy, but can apply them to its out of view mother. Further suppose that after a particular learning experience, the infant acquires the ability to apply *all* of these predicates to an out of view toy, but none to an out of view cat. Such a case would be most simply explained in terms of a non-conceptual way of thinking of the toy developing into a more conceptual way of thinking of the toy. It would be quite apposite to suggest that all of the predicates were non-conceptual prior to the learning experience, and they all increased in conceptuality, but only with respect to one object, not any others.

It will help to introduce some terminology. A *proto-concept* is a content constituent[13] of a content that is not a concept, nor a singular sense, nor is it composed completely of those. "Proto-concept", like "concept", will refer both to the actual constituent of the content and to the corresponding ability to entertain contents that contain that constituent.

From the preceding explication of what it is to be a concept, we can infer the following: if a content constituent is not a concept, then possessing it does not guarantee knowledge of what it is for something to be the semantic value of that constituent. That is, possessing a proto-concept does not entail knowing what it is for an arbitrary object to fall under the property to which the proto-concept refers (nor does it entail knowing what it is for an arbitrary object to be the referent of the proto-concept).

This may seem problematic at first. If possessing the proto-concept does not confer the knowledge of the semantic value as just elaborated, then why should one take it to have that semantic value in the first place? And if no other semantic value is available, that *does* correspond to the abilities conferred by possessing the proto-concept (a real possibility, since we are disallowing the possibility of gerrymandering into existence non-objective semantic values, an incoherent notion), then why consider the proto-concept to have a semantic value? And if it has no semantic value, then why consider it to be a constituent

[13] By restricting proto-concepts to being "content constituents" I intend to exclude constituents that are not concepts but that are also not themselves contents (such as objects in Russellian propositions).

of a content, or indeed a content at all? The notion of a non-conceptual constituent of content seems unintelligible.

The way out is to cut off the above line of reasoning with the first question. A general point I wish to make is that it is not always the case that one must have knowledge of the relation between a constituent and a semantic value in order to possess a content constituent with that semantic value. Although it is true that such knowledge is required in the case of conceptual constituents, the theory of content does not impose that requirement for content in general. To make sense of content for which this requirement does not hold, we need to make sense of a content constituent which has a semantic value, but the grasping of which does not require knowledge of what it is for something to be that constituent's semantic value. For the case of any predicate proto-concept F, we need to find a rationale for simultaneously holding a) that the semantic value of F is an objective property P, with its full range of application, and yet b) that possession of F, while it does confer the ability to entertain all contents that contain F as a constituent, does not confer the knowledge of what it is for an arbitrary object in the range of P to fall under P. Similarly, for the case of a subject proto-concept a, we need to make sense of simultaneously holding a) that the semantic value of a is an objective referent O, and yet b) that possession of a, while it does confer the ability to entertain all contents that contain a as a constituent, does not confer the knowledge of what it is for an arbitrary object to be the referent of a. Why we need to do *both* of these was explained in the last section: we must be prepared for the empirical possibilities that could require explanation via either non-conceptual subject contents or non-conceptual predicate contents.

Let F be a purported proto-predicate. What we would like to do is make sense of a gap between the contents that refer to objects that F can be true of (the *range of application* of F) and the subject concepts that F can be predicationally combined with (a rough notion of the *range of predication* of F). If this can be done, then we would have made sense of the idea of a proto-conceptual predicate: a content F the possession of which does not require knowledge of what it is for something to be the semantic value of F.

It will help to give more precise definitions of the two ranges mentioned above:

> **Range of Application:** The range of application of a predicate content F is the set of subject contents of whose referents F may hold or not hold.

The idea of a range of application is widely recognized as a prerequisite for stating the Generality Constraint (although it is usually stated indirectly via the level of content rather than directly on the level of reference as I have done here). When first stating the Constraint, Evans notes that "a proviso about the categorial [*sic*] appropriateness of the predicates to the subjects" is required [Evans, 1982, p 101 fn]. Likewise, Cussins states the Constraint thusly:

> An organism does not possess a concept *a* of an object unless it can think *a is F*, *a is G*, and so on for all of the concepts *F*, *G*, ... of properties which it possesses (*and which are not semantically anomalous in combination with *a**).[Cussins, 1990, p 417, emphasis added]

Similarly, Peacocke appeals to the range of application, although he expresses it in terms of a referential-level *range of significance*. Recall his statement of the Generality Constraint from section 3.3.2:

> **Generality Constraint:** If a thinker can entertain the thought Fa and also possesses the singular mode of presentation b, which refers to something in *the range of objects of which the concept F can be true or false*, then the thinker has the conceptual capacity for propositional attitudes containing the content Fb.
>
> ...It will be convenient to label the range of objects of which a given concept is true or false its "range of significance" [Peacocke, 1993, p 42, emphasis added].

The range of application is distinct from the range of predication:

> **Range of predication:** The range of predication of a predicate content F is the set of all singular senses a that are such that, if a subject possesses F, the subject's possessing a ipso facto enables the subject to entertain the thought Fa.

Now suppose that the range of predication of some content F is a subset of the range of application of F. It follows that there exists a subject content a which refers to an object o such that F can be true or false of o, and yet a subject possessing both F and a would not thereby be guaranteed the ability to entertain the thought Fa. That is, in such a case, the subject can think thoughts involving F, without knowing what it is for an arbitrary object (in the range of application of F) to fall under the semantic value of F. That is, the subject would not know what it is for something to be the semantic value of F. The Generality Constraint would therefore not apply to such a content, and thus it is not a conceptual content.

A non-General subject content is simply this: a content a for which there is an F such that a is not in F's range of predication, yet a refers to something in the range of application of F's referent. The account could have been given from the perspective of subject contents (perhaps using the correlative notions of range of applicability and range of predicability) instead, with the predicational case being the derivative case, but that is merely a matter of notation and style.

Before, in discussing the Possession Principle (section 3.3.3), I explained that making the conceptual/non-conceptual distinction requires that there be a way for the predicational combination of F and a to fail to be a content in some way other than being a semantic anomaly. I can now make good my promise to explain how this can be. Suppose there is a gap between F's range of application and range of predication. That is, suppose that a is within F's range of application, and yet is not in F's range of predication. If the Possession Principle is correct, then it can be concluded that there is no such content Fa. For if there were, the fact that a is in F's range of application, together with the possession of both F and a, would imply the ability to entertain Fa. Yet by stipulation they do not. Therefore, we have another kind of anomaly: even if F and a are not semantically anomalous, it may be that F is *predicationally anomalous* in combination with a.

3.5 Judgement and Generality

Let us take stock. I have explained what a constituent of a content is by explaining what constituent structure is: it is that which best explains the cognitive significance, reference, and entertainability relations of a content. I have argued that one should not make a distinction between conceptual content and non-conceptual content on the grounds that they have distinct types of reference. I also gave reasons for rejecting a conceptual/non-conceptual distinction that is stipulated with respect to the Generality Constraint. That left cognitive significance as the primary factor in making the conceptual/non-conceptual distinction. I have suggested (chapter 2, section 2.1) that concepts should be stipulated to be the constituents of contents which are individuated with respect to their relations to judgement alone. I now justify this stipulation by showing how it can explain the Generality Constraint. Then simple contraposition gives us the conclusion that content which is not General (not combinable with all contents of the appropriate semantic category) is non-conceptual; that is, content which is individuated with respect to more than its role in judgement.

The derivation of the Generality Constraint is as follows. First, let us accept the stipulation that conceptual content is individuated with respect to role in judgement alone. An important insight is this: what individuates a content also fixes its semantic value.[14] Thus, a conceptual content's role in judgement fixes its semantic value.

This claim is implicitly supported by Peacocke's views on concepts and reference:

> As a matter of principle, the level of reference is inextricably involved with concepts, as understood here. Concepts are individuated by their possession conditions; the possession conditions mention judgements of certain contents containing the concepts; judgement necessarily has truth as one of its aims; and the truth of a content depends on the references of its conceptual constituents. It would be wrong, then, to regard the referential relations in which concepts stand as grafted onto a structure of concepts that can be elucidated without any reference to reference. Referential relations are implicated in the very nature of judgement and belief [Peacocke, 1993, p 17].

That a concept's role in judgement fixes its semantic value can be explained as the result of two factors. First, there is what Peacocke states above: judgement has truth as its aim. He says more about this elsewhere:

> The attitude of judgement is well-suited to occupy a central position in an account of content. This is so because truth is internal to judgement in a way in which it is not internal to many other attitudes. The point is not that to judge p is to judge that it is true that p. For equally to hope that p is to hope that it is true that p, and to fear that p is to fear that it is true that p; but hope and fear should not occupy a central position in a theory of thought. The point is rather that truth is one of the *aims* of judgement. A thinker aims to make this the case: that he judges that p only if it is true that p... On the notion of content used throughout this essay, what is required for a given content to be true is always intrinsic to that content: it is something constitutive of that content's identity. Judgement aims at truth of the content judged. So in learning what is necessarily involved in judging a given content, we can learn

[14] More precisely, it fixes the content's contribution toward determining a semantic value. The way the world is may contribute to the determination of a content's semantic value. This might make one think that what I say in the rest of this section implies that if the way the world is partially determines a content's semantic value, then that content must be non-conceptual. But it does not.

> something about the nature of that particular content itself [Peacocke, 1986, p 46-7, emphasis in the original].

The second factor is that the notion of semantic value is a truth-based notion. That is, there is nothing more to the semantic value of a content than the way that it contributes to the truth of contents which contain it as a constituent. Thus, since judgement aims at truth, a concept's role in judgement determines its semantic value.[15] Since in the work just cited, Peacocke's concern with "thought" is a concern with conceptual content only, if my stipulation concerning the nature of conceptual content is right, it should be that Peacocke is concerned only with contents that are individuated with respect to judgement alone. So he is; thus he says: "hope and fear should not occupy a central position in a theory of thought". As I am concerned with what might be thought of as a broader notion of thought (or alternatively, a broader notion of content that includes more than the content of thoughts), I am thereby concerned with a theory of content in which judgement is not the only cognitive activity that is given a central place.

Now we can show that a content individuated solely by its relations to judgement meets the Generality Constraint. Suppose that a predicate content so individuated did *not* meet the Constraint. From the previous section, we know that that would imply that there is a gap between that content's range of application and its range of predication. But this is impossible: the predicate's relations in judgement alone determine the range of application, via determining the concept's semantic value, which is in turn determined via those relations individuating the concept as a whole, including its range of predication. If there were to be some gap, it would have to be that the concept is individuated with respect to relations in judgement that do not determine a range of predication that is equal to the range of application. And that gap cannot be by virtue of there being elements "missing" from the range of application that are in the range of predication; that would be to dabble in the idea of sub-objective reference, a possibility which I have already argued against (see section 3.3.4). Therefore, it must be that there are elements in the range of

[15] Now a question, that was raised in a footnote concerning step 9) of Peacocke's derivation of the Generality Constraint (section 3.3.3), can be answered. It is because of the essential connection between truth and semantic value that for a subject to know what it is for a particular thing, B, to be the referent (semantic value) of b, the subject must know what it is for an arbitrary object to be the referent of b.

application which are not in the range of predication. That is, the judgemental relations taken alone would fix a range of predication which is less than the range of application. But then, given the connection between judgement and truth, so also would those relations not fix the full range of application; something else would have to take up the slack in individuation so that the full range of application is fixed, else a semantic value would not be determined. That is, the concept would have to be individuated with respect to more than just its role in judgement. But this contradicts the primary hypothesis that we are dealing with a conceptual content. So we must reject the secondary premise: it must be that the concept is General after all.

In summary:

1. Conceptual content is defined as content that is individuated with respect to judgemental relations alone.

2. If one individuates a predicational content F on the basis of judgemental relations alone, then a unique semantic value is determined only when the individuation employs all the judgemental relations between all the contents that are formed by combining F with each of the modes of presentation in the range of application of F.

3. If a content is individuated in the above manner, there is no possibility, for any modes of presentation a in the range of application of F, of Fa being predicationally anomalous.

4. If Fa is not predicationally anomalous, then anyone who possesses F and a and grasps the mode of combination will not be conceptually barred from grasping Fa.

More light can now be shed on Peacocke's derivation of the Generality Constraint. We left that derivation with the conclusion that Generality followed from concepts being a particular kind of content: content which in grasping, one thereby knows what it is for something to be its semantic value. If the foregoing is right, it must be that a content's being individuated solely by relations in judgement makes it such that when one grasps it, one thereby knows what it is for something to be the semantic value of it. This again is explained by the joint factors of 1) the truth-directedness of judgement and 2) that contributions to truth are constitutive of semantic value.

3.6 Conceptuality and objectuality

In making the preceding argument, I established a lemma: any content which fails to meet the Generality Constraint must be individuated with respect to something other than judgement alone. In chapter 2, I argued that the Example requires explanation in terms of sub-objectual content. To explain why the Example demands non-conceptual content, I now only need to show the following: that non-objectual content is not General.

In chapter 2, I argued that to think of something objectively is to think of it objectually. Therefore, non-objectual content (which the Example requires) is non-objective. I have made it clear that this non-objectivity is not manifested in a lack of determinate truth conditions, nor is it manifested in the content having, per impossibile, a sub-objective referent. Rather, it is manifested in the conditions under which the content may be entertained.

Support for the claim that objectivity requires Generality can be found in Strawson [Strawson, 1959]. The structure of his position, as argued in the first chapter of *Individuals*, is:

1. Reference to objective particulars requires identifying them;
2. Identifying an objective particular requires locating it within a unified, systematic framework of particulars;
3. Locating a particular in a unified framework requires the ability to re-identify that particular;
4. Re-identification of a particular requires the ability to conceive of existence unperceived.

The first point, understood in an unrestricted sense, would be at odds with my conception of non-conceptual content. On my account, one precisely *can* refer to objective particulars (although not *as* such) without being able to locate them in a unified framework, to re-identify them, or conceive of them existing unperceived. For something to be a referent, it must be objective.

But Strawson means something more restricted when he speaks of referring to objective particulars. He argues for 1) thusly:

> It is not merely a happy accident that we are often able, as speakers and hearers, to identify the particulars which enter into our discourse. That it should be possible to identify particulars of a given type seems a necessary condition of the inclusion of that type in our ontology. For what could we mean by claiming to acknowledge the existence of a class of particular things and to talk to each other about members of this class, if we qualified the claim by adding that it was in principle impossible for any one of us to make any other of us understand which member, or members of this class he was at any time talking about?[Strawson, 1959, p 16]

This constraint on conceptual thought is echoed by Evans [Evans, 1982]. Evans advocates a fundamental constraint on thought which he calls "Russell's Principle": "a subject cannot make a judgement about something unless he knows which object his judgement is about" [Evans, 1982, p 89]. Like Strawson, Evans is assuming, in insisting on this identificational requirement, the case of conceptual content.

With respect to this conceptual mode of reference, I agree with both Strawson and Evans: reference *of this sort* requires an identification of the referent. When Strawson considers "reference to objective particulars", he is considering a more restricted notion of reference than I do when I claim that even non-conceptual contents have objective referents. I am using "objective" to modify only the referent, while Strawson uses "objective" in a way that also constrains the mode of referring. That is, Strawson means a mode of referring which is conceptual, a mode of referring in which it is the case that one can only refer to something that is in one's *ontology*, a mode in which one can only refer to something of a particular class if one also *acknowledges* the existence of of that class. Similar reasoning applies to Evans' case: he is concerned with a notion of reference that is more restricted than the mere possession of a semantic value, with which I am concerned. Evans explicitly states that his defense of Russell's Principle relies on the assumption that we are dealing with a notion of thought such that "in order for a subject to be credited with the thought that p, he must know what it is for it to be the case that

p" [Evans, 1982, p 105]. From the discussion the first sections of this chapter, we can see that both Strawson and Evans are restricting themselves to conceptual content: content made up of constituents to possess which a subject must have knowledge of what it is for something to be the semantic value of each constituent. Thus, their arguments show that it is conceptual content which requires locating a particular in a unified framework (Generality), which in itself requires the ability to re-identify particulars and therefore the ability to conceive of existence unperceived (objectuality).

In explication of the framework mentioned in the second claim, Strawson says:

> Yet it cannot be denied that each of us is, at any moment, is in possession of such a framework – a unified framework of knowledge of particulars, in which we ourselves and, usually, our immediate surroundings have their place, and of which each element is uniquely related to every other and hence to ourselves and our surroundings.[Strawson, 1959, p 24]

And later:

> We operate with the scheme of a single, unified spatio-temporal system. The system is unified in this sense. Of things which it makes sense to inquire about the spatial position, we think it is always significant not only to ask how any two such things are spatially related at any one time, the same for each, but also to inquire about the spatial relations of any one thing at any moment of its history to any other thing at any moment of its history, when the moments may be different. Thus we say: A is now in just the place where B was a thousand years ago. We have, then, the idea of a system of elements every one of which can be both spatially and temporally related to every other.[Strawson, 1959, p 31]

Thus Strawson maintains, as does Evans, that at least in the case of thinking about spatio-temporal particulars, conceptual thought is manifested in the possession and maintenance of a unified conceptual framework within which the subject can locate, and thus relate to any other arbitrary object of thought, the particular being thought about. That is, conceptuality requires Generality. This is in effect an explanation of the Generality Constraint: conceptual content makes use of such a unified framework in order to locate the particulars and properties thought about within a relational space involving all other relevant objects and properties. Thus we have reached step 2) in Strawson's position, as outlined above.

Evans develops this idea of a unifying framework that underlies our capacity for conceptual thought, arguing that, at least in the case of fundamental thought about places, it could be realized in the maintenance and use of a cognitive map [Evans, 1982, p 151]. Cussins furthers the development, taking the cognitive map, a computational structure used in spatial navigation, to be the model of a unifying structure for all domains of thought [Cussins, 1990]. I further develop these suggestions in detail in chapters 7 and 8.

Strawson's argument for 3) appeals to our inability to encompass the totality of the world, even that portion of the world with which we are concerned, at any given moment:

> Why are criteria of reindentification necessary to our operating the scheme of a single unified spatio-temporal framework for referential identification? The necessity may be brought out in the following way... [W]e do not use a different scheme, a different framework, on each occasion. It is the essence of the matter that we use the same framework on different occasions... We cannot attach one occasion to another unless, from occasion to occasion, we can reidentify elements common to different occasions. [Strawson, 1959, p 32]

But reidentification requires the ability to conceive of existence unperceived; Strawson argues for 4):

> Our methods, or criteria, of reidentification must allow for such facts as these: that the field of our observation is limited; that we go to sleep; that we move. That is to say, they must allow for the facts that we cannot at any moment observe the whole of the spatial framework we use, that there is no part of it that we can observe continuously, and that we ourselves do not occupy a fixed position within it [Strawson, 1959, p 32].[16]
>
> ...[A] *condition* of our having this conceptual scheme is the unquestioning acceptance of particular-identity in at least some cases of non-continuous observation [Strawson, 1959, p 35].

At last, Strawson has given us the conclusion we sought: conceptual thought is objectual thought. Thus, the content that we need to account for the non-objectual data in the Example must be non-conceptual content.

[16] Although the way Strawson puts the point makes it look as if it is merely a contingent fact that we cannot encompass the totality of the world at any given moment, I think that this limitation is not only implied by, but is constitutive of, what it is to be a subject. To be a subject is to something for which the term "subjective" can apply; it is to have a point of view. Yet the idea of having the totality of the world always available to one is not an idea of a point of view, but an idea of having no point of view (see chapter 7, section 7.4.1).

3.7 The Example revisited

Although it has already been shown that the non-objectuality of the Example demands an account in terms of non-conceptual content, some more can be said about this.

If the argument in this chapter is right, then it must be the case that the contents of the infant in the Example are individuated with respect to more than just their roles in judgement. But it would be a mistake to confuse the fact that a content is individuated with respect to more than its role in judgement, with the fact that only connections in judgement are used as evidence for the ascription of the content in a particular case. For even if the latter fact is the case, it may still be the case that the content so ascribed is individuated with respect to more than judgement. Even so, there is no need to invoke this insight, since it seems very likely that both the evidence for ascription of, and the individuation of, the contents involved in the Example will appeal to relations to contents other than judgement relations. For example, the lack of Generality concerning an infant's way of thinking of an object means that that content's connections in judgement to other contents are not sufficient to determine an objective semantic value. But the content's *developmental* connection to a conceptual content, which does have a clear referent, may be sufficient to determine a referent for the non-conceptual content. These developmental relations which partially individuate a content c might include something like:

$$
\begin{array}{l}
< C_1, e_1, c, F_1 > \\
< C_1, e_2, c, F_2 > \\
< C_2, e_1, c, F_3 > \\
< C_2, e_2, c, F_4 > \\
. \\
. \\
.
\end{array}
$$

where $< C_i, e_j, c, F_k >$ represents the relation:

> If the agent takes the belief attitude to the contents in C_i, and subsequently has experiences e_j then, in normal conditions, c will develop into the conceptual content F_k.

However, it may be that the referents of the infant's contents in the Example can be determined in a more straightforward manner. In chapter 2, section 2.4.5, recent data in developmental psychology are mentioned, data which some might take to undermine my insistence that the Example involves pre-objectivity. I pointed out that even if the infant's behaviour respects various objectivity constraints in one domain, this may not be sufficient for the ascription of an objectual content in that domain, let alone in other domains involving the same object. However, not only do these recent data not argue against a non-conceptual approach, they actually may assist it. Consider again Baillargeon et al's experiment: the infant is startled when a rotating screen passes through a region where an unperceived toy should be. Inasmuch as these data suggest so strongly to Baillargeon et al that the infant has an objectual concept of a toy, they can instead serve the role of justifying the ascription of a proto-concept with the toy as its referent, despite the fact that the proto-concept fails to meet other constraints associated with a conceptual way of thinking of the toy.

Thus we have an answer to the question raised in section 3.4: If possessing a proto-concept does not confer the knowledge of what it is for something to be its semantic value, then why should one take the proto-concept to have that semantic value in the first place? The answer is a direct application of the general findings of this chapter: it may be that a content's relations in judgement are insufficient to constitute the knowledge of the semantic value, and yet when these judgement relations are supplemented with others, a semantic value is determined. This shows non-conceptual content to be externalist in the sense that what we refer to outstrips our conceptual knowledge of it.

In section 2.3 of chapter 2, I stated that one of the advantages of the Example is that it involves transitions which increase the conceptuality of an infant's interactions with its environment, a dynamic which also requires a non-conceptual understanding. The idea of increasing the conceptuality of a content constituent requires a notion of conceptuality that admits of degrees. This notion needs elucidation, which can now be provided. That the conceptuality of a content admits of degrees is a consequence of the Generality Constraint being essential to conceptual content. Because a constituent failing to meet the Generality Constraint can be understood in terms of a gap between that constituent's range of application and its range of predication (section 3.4), a partial ordering can be imposed on the gap (the difference between the two sets), and this ordering can be used to give meaning to statements such as "content constituent X is more conceptual than constituent Y": X will be more general than Y if the gap between its ranges is smaller than the gap between Y's ranges.

Thus, there is not just the simple distinction between one level of content which is non-conceptual, and one level which is conceptual. Rather, there are numerous degrees of conceptuality/non-conceptuality. Chapter 8 is an extended investigation into the notion of degrees of conceptuality; specifically, I propose quantitative measures for rating the Generality and hence conceptuality of the representations, and thus contents, of a artificial cognitive map for spatial navigation.

Consonant with the gradedness of conceptuality and the link that I have argued exists between conceptuality and objectuality, this gradedness has also been observed with respect to objectivity and objectuality. The notion of graded abilities of infants is at least as old as Piaget; but what is being discovered now is that the development of objectivity has a richer texture, worthy of its own intentional account, than even Piaget imagined. First, recall Harris' observation, from section 2.4.4 of chapter 2, concerning the $A\overline{B}$ data:

> These very orderly data create a problem for any purely conceptual or cognitive interpretation. Consider, for example, the idea... that the infant does not appreciate that an object can be in only one place at a time, and fails to rule out A as a possible hiding place. Are we to say that the 8-month-old baby can rule out A for 3 seconds but no longer, the 9-month-old baby for 5 seconds but no longer, and so forth? If the baby has come to understand that an object can only be in one place, why does it not apply this knowledge to delays of any length? [Harris, 1989, p 115]

This is a clear example of a gradual, incremental increase in the degree to which the infant meets the objectuality constraints discussed in that chapter. Also, Hood and Willatts [Hood and Willatts, 1986] have observed that some (i.e., five-month-old) infants that lack our full notion of object-permanence can represent objects that are unperceived, but only under certain conditions (e.g., in the dark). This is just one instance of an intermediate stage between no notion of object permanence at all, and the full, objective notion which we take ourselves to possess. The development of objectivity is not one single, boot-strapping, mechanistic jump from a non-intentional system to a fully objectual one; rather, it is a complex intentional phenomenon which itself requires explanation, and which constrains the possible explanations of the relatively objective form of cognition that it yields. The stages in this development cannot be understood conceptually; the differences between the infants of one age and those of another are not differences in accepted conceptual contents, or in the possession of a set of concepts. Rather, the differences are internal to the proto-concepts they are employing in such cases. An explanation of the development of conceptuality cannot itself appeal to the concepts it is attempting to explain.

This makes good another promise from chapter 2. There the soft line for establishing the non-objectuality of the Example was introduced. The advantage of the soft line is that it relies on weaker, and therefore less contentious, premises. It admits that perhaps an objectual story can be told for the infant in the Example, but it is inferior to a non-objectual story. There are three stages to the soft line. First, one must make sense of non-objectual content; this was done in section 2.4.4 of chapter 2. Then one makes sense of an ordering of less objectual and more objectual ways of thinking. This has been done: the connection between conceptuality and objectivity on the one hand, and the gradedness of conceptuality on the other, gives us a notion of degrees of objectuality (a notion that is given extensive illustration in chapter 8). Finally, one applies a principle analogous to Lloyd Morgan's Canon[17] to state that the more basic account is to be preferred to the more complex, objectual one.

What was said about the Example in chapter 2 connects with the results of this chapter in section 3.6. The argument there relies heavily upon Strawson's and Evans' arguments which connect Generality and objectivity, arguments which are at times obscure. To further clarify the connection, I provide, in chapters 7 and 8, an extended illustration in the form of a simulated agent which navigates an artificial world. There it is made clear how the non-Generality of the agent's ways of thinking of places results in an inability to conceive of places as existing independently of the agent.

[17] "*In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale*" [Lloyd Morgan, 1894, p 55, emphasis in original].

3.8 Two objections

An extended objection to the notion of non-conceptual content is addressed in chapter 4. Before moving on to that, I consider two objections to the notion of non-conceptual content, one based on a requirement for determinate truth-conditions, and one based on the Generality Constraint's reference to conceptual capacities.

3.8.1 Non-conceptual content and truth

All along I have been employing a notion of non-conceptual content which has determinate truth-conditions. That is, whole non-conceptual contents may be true or false. One way of objecting to this notion of non-conceptual content, then, is to claim that a content can have determinate truth-conditions only if it is conceptual.

Cussins argues for this claim. As a proponent of non-conceptual content, his intention is not to argue against the possibility of non-conceptual content in general. Rather, he takes it that such content does not have determinate truth-conditions; it must have norms other than strict truth or falsity. But his argument reflects a general line of thinking that, together with the insistence that content must necessarily have determinate truth-conditions, leads some to believe that non-conceptual content is incoherent.

His argument (all premises are from [Cussins, 1990, pp 384-387] unless stated otherwise) is:

1. Abbreviate "content with determinate truth-conditions" to "α content".

2. The interpretation of a cognitive occurrence as having α content depends on specifying the content by means of a representational state (e.g., a linguistic expression).

3. All representational states contain, either implicitly or explicitly, indexical or demonstrative elements [Cussins, 1990, p 391, fn 45].

4. An α content may have no unrelativized indexical or demonstrative elements (implicit).

5. Indexical or demonstrative elements (that are contained in a representational state

being used to specify the α content of a cognitive occurrence involving a behaviour) are relativized with respect to the concepts of a task-domain for that behaviour.

6. "A task domain [for a behaviour] is a bounded domain of the world which is taken as already registered into a given organization of a set of objects, properties or situations, which contains no privileged point or points of view, and with respect to which the behaviour is to be evaluated" (stipulated).

7. Therefore, "the interpretation of a cognitive occurrence as having α content depends on specifying the content by means of concepts of a task-domain".

8. "An organism can only grasp an α content if it grasps its truth-conditions (or its contribution to the truth conditions of contents containing it)".

9. Therefore, an organism which grasps an α content "must know what the (relevant part of the) t[ask]-domain of the content is. But a t[ask]-domain (unlike the world) is essentially conceptually structured, so there is no way of knowing what the t[ask]-domain of a content is without possessing the concepts in terms of which the t[ask]-domain is structured. Hence, possession of an α content requires possession of the concepts in terms of which it is canonically specified."

10. Conceptual content is content the possession of which requires the organism to possess the concepts in terms of which it is canonically specified (see section 3.3.1).

11. Therefore, α content is conceptual content.

The illicit step in the argument which is most germane to the issues discussed so far occurs at 8). Cussins offers no support for this premise, but rather relies on the traditional theory of content. However, as Cussins well knows, this theory has not recognized the existence of non-conceptual content, so many of its findings have implicitly assumed the case of conceptual content. If what I have said so far is right, then 8) does not hold for content with determinate truth-conditions in general, but only for conceptual content. Indeed, 8) follows from that aspect of conceptual content that explains its Generality, as we saw in the discussion of Peacocke's derivation of the Generality Constraint (section

3.3.3). Knowing what it is for something to be the semantic value of a concept implies grasping that concept's contribution to the truth conditions of contents which contain it. But it is only in the case of conceptual content that one must have this knowledge in order to possess the content in question. Therefore no contradiction is forced if one assumes that there is non-conceptual content with determinate truth-conditions.

Note that the argument, once it is seen as applying to conceptual content only, succeeds in showing one of the things that I had set out to show. In section 3.2, I acknowledged a desideratum for a stipulative notion of conceptual content that is distinct from conventional definitions: that it agree with those conventional definitions, and explain why conceptual content has the properties used in those definitions, even if it is not to be defined in term of them. This explanation was provided for the Generality Constraint, and Cussins' argument does the same for the other essential property of conceptual content, discussed in 3.3.1: it *explains* why possessing a conceptual content requires possession of the concepts used to canonically specify it. The same insight also explains why non-conceptual content does not require those who grasp it to possess the concepts used in specifying the content: grasping a non-conceptual content does not require one to grasp the truth-conditions of the content.[18]

3.8.2 Conceptual capacities

Recall one more time Peacocke's statement of the Constraint:

> **Generality Constraint:** If a thinker can entertain the thought Fa and also possesses the singular mode of presentation b, which refers to something in the range of objects of which the concept F can be true or false, then the thinker has the *conceptual capacity* for propositional attitudes containing the content Fb.[Peacocke, 1993, p 42, emphasis added]

He explicates what he means by "conceptual capacity":

> By speaking of a "conceptual capacity" for attitudes with the content Fb, I do not mean merely that the thinker possesses the conceptual constituents

[18] Admittedly, there are some steps in Cussins' argument that need support, even if one does restrict it to conceptual contents. For example, in chapter 5 I explore the possibility of specifications of content that do not proceed by offering a representational state that has the same content as the one to be specified. I thus do not accept Cussins' 2) as obvious.

> of Fb. Nor by "conceptual capacity" do I mean that the thinker will easily entertain the thought that Fb. For certain contents, such mechanisms as self-deception, repression and the like may prevent the thought from being so much as entertained. There may also be preventing factors at the level of hardware. A thinker might really have a language of thought, and it might be that attempts to concatenate his Mentalese symbols for the concepts F and b produce strange chemical reactions that prevent him from entertaining the thought Fb. What I do mean by a conceptual capacity for attitudes with the content Fb is this: The thinker is in a position to know what it is for the thought Fb to be true. That is, if there is some block to the thinker's attaining states with the content Fb, what is missing is not any knowledge about concepts, nor any conceptual capacity [Peacocke, 1993, pp 42-43]

On reading these remarks, one might be inclined to reason as follows. All putative data that seem to demand the ascription of a content that does not meet the Generality Constraint can in fact be explained instead by ascribing a conceptual content and attributing any apparent inability to recombine the constituents of that conceptual content to something other than the absence of a conceptual capacity. In the case of the Example, for instance, it could be that the infant does indeed possess the conceptual capacity to entertain the thought *Behind-me(the toy)*. The failure of the infant to actually entertain that thought, or to exhibit any behaviour that provides direct evidence for the capacity to entertain that thought, could be attributed to the kind of failures that Peacocke mentions. The objection is not that this interpretation will be possible in some cases; that claim is entirely consistent with a notion of non-conceptual content. Rather, the objection is that there will always be an interpretation of the data that attributes any apparent lack of Generality to a failure not involving a lack of the relevant conceptual capacities.

The problem with this objection is that it seems to be appealing to a Cartesian notion of concepts, a view which I criticized in chapter 2, section 2.4.4. That is, it employs a notion of concept possession that is completely independent of empirical, contingent fact. No matter what the details of the subject's physical makeup, any failure to entertain the content will not, a priori, be attributable to the absence of a conceptual capacity. One must be careful, however, how one uses this aspect of the objection against itself. It will not do, for instance, to counter the objection with the following:

> In any case of apparent non-Generality, whether or not the failures that Peacocke mentions (repression, hardware failures, etc) are present is an empirical matter. Thus there is always the empirical possibility that none of these

> failures are present, leaving open only one remaining explanation for the non-Generality: the absence of conceptual capacity, and thus confirmation that the data involve a content for which the Generality Constraint does not hold.

This is inadequate, because given the very naturalism that this reply is assuming, there will always be some non-intentional explanation of the behaviour in question which the opponent of non-conceptual content can offer instead of an explanation involving an absence of a conceptual capacity.

What needs to be done is to provide the conditions under which an explanation of non-General behaviour is an explanation in terms of the lack of a conceptual capacity, rather than trying to rule such explanations in or out, a priori. The reply to the objection should employ a very familiar maxim: a higher level explanation can be warranted, even in the face of lower-level explanations, as long as it is simpler, or provides generalizations not expressible at the lower level, etc. Thus, even if there is an explanation of non-General behaviour in terms of hardware faults, those same faults may underwrite a simpler explanation in terms of the absence of a conceptual capacity.

An illustration of how this could be possible will be of use. For example, the infant in the Example might seem to be able to think *Red(the toy)* and *Behind-me(mother)*, yet it does not appear to be able to entertain the content *Behind-me(the toy)*. No doubt there will be some non-intentional (although perhaps very complex) explanation of why the infant exhibits the behaviour that warrants the ascription of the thoughts *Red(the toy)* and *Behind-me(mother)*, but cannot behave in a way that warrants the ascription of the thought *Behind-me(the toy)*. But suppose also that the infant cannot think *Behind-me(x)* for any objects x it can think of in other contexts, but which do not make sounds. Similarly, suppose the infant can think *Behind-me(x)* for all objects x it can think of in other contexts, that do make sounds. And further suppose that this pattern holds for all predicates that, like *Behind-me*, typically require objects that they are true of to be visually unperceived, but audibly perceivable. Then one explanation becomes irresistible: the particular failure to entertain the thought *Behind-me(the toy)* is a consequence of an absence of a general conceptual capacity: the ability to conceive of objects as existing unperceived. Thus we have an illustration of how the empirical facts can form a pattern that requires one to attribute a failure of Generality to a lack of a conceptual capacity.

Such a case therefore requires an explanation involving non-conceptual content.

CHAPTER 4

A Neo-Kantian Critique

One kind of worry concerning not just the utility, but the coherence, of the notion of non-conceptual content is expressed in the slogan: "contents without concepts are blind". When first encountering the notion of non-conceptual content, worrying this worry is a common reaction. I will dispel this Kantian angst, by responding to its latest manifestation, in the John Locke lectures given by John McDowell [McDowell, 1994b].

McDowell's wording of the slogan – "*experiences* without concepts are blind" – is somewhat closer to the Kantian original (which itself talks of intuitions, not contents). In saying that experiences without concepts are blind, McDowell does not mean that there is a kind of experience, blind experience, which is non-conceptual. Rather, he intends "blind experience" to be an illuminating oxymoron. There is no experience which has non-conceptual content: all experience is conceptual. He offers this conclusion as the only way to avoid an unacceptable oscillation between two extreme views: on one hand an empty coherentism, which denies any rational connection between concepts and experience; on the other, a blind Mythology of the Given, which posits experience as rationally linked to, yet distinct from, the conceptual. McDowell's objection to the former position is easy to state: it cannot make sense of the fact that our experiences are about the world. But I will take more care in stating McDowell's objections to the latter position, because it is this position that he takes to be the one (unintentionally) occupied by advocates of non-conceptual content. And it is, of course, a position which he argues to be untenable.

4.1 McDowell's target

McDowell's explicit target is a particular kind of non-conceptual content, the kind that, e.g., Peacocke employs to provide the possession conditions for observational concepts

[Peacocke, 1993]. Whereas that notion focusses on the, as it were, horizontal links between non-conceptual content and concepts, this thesis has employed a notion of non-conceptual content that focusses on the vertical links; development over time, rather than transduction at a time, from the non-conceptual to the conceptual (and vice versa). Thus it is tempting to dismiss McDowell's discussion as irrelevant, given that his arguments are attacking that particular, different, horizontal notion of non-conceptual content. Specifically, McDowell criticizes the role that non-conceptual content can play in *justifying* the applications of our observational concepts, whereas I have been arguing for a notion of non-conceptual content that can perform a different explanatory task: the explanation of action and behaviour. Vertical non-conceptual content can be the object of the attitudes, and of judgement. In contrast, horizontal non-conceptual content is outside the sphere of the attitudes and judgement, but is normatively linked to the contents within that sphere. Given such fundamental differences, it might seem that I could accept McDowell's conclusions for horizontal non-conceptual content and still maintain that vertical non-conceptual content is immune.

However, I cannot so isolate myself from McDowell's concerns. His arguments against the ability for non-conceptual contents to provide rational justifications for attitudes toward conceptual contents also tell (a fortiori) against the possibility of such links between one non-conceptual content and another. The explanations I wish to make room for explain action by providing personal-level, albeit non-conceptual, contents that rationalize the action to be explained. Thus, in this explanatory scheme, vertical non-conceptual content can be the objects of attitudes such as belief and desire, and can constitute reasons; and it is against the possibility of extra-conceptual reasons that McDowell argues.[1]

Putting the point that way, it looks as if the success of McDowell's arguments would be adverse only to the particular application I have in mind for vertical non-conceptual content. But more is at stake than that; vertical non-conceptual contents are (partly)

[1] McDowell's strategy is to show that non-conceptual contents cannot provide reasons, and then appeal to the fact that only by having reasons can our conceptual empirical judgements be non-empty, be about the world. The fact that McDowell does not provide arguments for the second stage of his strategy weakens his project as a whole, but it does not weaken the threat, provided by his arguments for the first stage, to the notion of non-conceptual content I am putting forward.

individuated by their rational, reason-providing connections to other contents. If such links are impossible, then there can indeed be no such thing as non-conceptual content at all, be it horizontal or vertical.

There is another line of reasoning that might lead one to believe that vertical non-conceptual content is acceptable to McDowell. It might be thought that since such contents can be the objects of the attitudes, they are, by McDowell's lights, conceptual, and thus unobjectionable. I show, in section 2 below, that this way of understanding McDowell is incorrect. McDowell's notion of the conceptual no doubt does include being the objects of the attitudes, but it includes much else besides. In particular, having an experience with conceptual content implies a grasp of the subject/world distinction and the concept of truth, which I have argued are not implied by the entertaining of certain non-conceptual contents. And these other features play a crucial role for McDowell: in effect, he argues that there can be no contents which do not have these implications. Thus, McDowell will not be able to accept vertical non-conceptual content, despite the fact that its capability of being the object of the attitudes makes it crucially different from horizontal content.

Finally, although McDowell's expression of the arguments always assumes a conceptual inner self, the arguments themselves are general enough to apply to the putative non-conceptual contents of an organism with no concepts at all.

4.2 Non-conceptual experience

For McDowell, (the contents of) experience must justify our empirical thinking (on pain of emptiness). Furthermore, there must also be external constraints on our conceptually driven freedom, our spontaneity: we cannot experience whatever we want to experience. The Myth of the Given is the idea that a notion of extra-conceptual experience can play this double role: being experiential, it is rationally linked to the conceptual; but being extra-conceptual, it is external. But, McDowell argues, if experience is extended beyond the sphere of the conceptual, then its very externality means that it cannot provide the justifications required. Because it is outside of our spontaneity, outside of our control, then external experience can offer us exculpations: if we are not in control, we cannot be blamed. But exculpations are the most it can offer; for if we cannot be blamed, neither can we be justified in believing the inner contents to which the extra-conceptual impingements give rise.[2]

This reasoning rests on a confusion. It is true that in as much as we cannot be blamed for our experiences, we also cannot be justified in having them. But it is not for our experiences that we seek justifications; but rather our thoughts and our thinking. It is thought that can be justified by experience, and it is thought for which we can be responsible.

McDowell acknowledges this objection (p 53, fn), and addresses it in the postscript to his lecture on non-conceptual content. Earlier, McDowell asserts that experiences must justify thoughts, but they can only do that if they have conceptual content. In the postscript he admits that there is a notion of justification under which it can be said that non-conceptual experiences justify, and are thus rationally linked to, beliefs. But he takes these to be justifications only in a manner of speaking. They are not the kind of justifications that can constitute a *subject's reasons* for believing. He makes the distinction with an analogy. The movements a skilled cyclist makes while rounding curves might be rational in that they are:

> ...suited to the end of staying balanced while making progress on the desired trajectory... But this is not to give the cyclist's reasons for making those movements... Why would it not be similar with experiences and judgement, if experiences had the non-conceptual content Peacocke says they have? [McDowell, 1994b, p 163].

The analogy is not all that satisfying. The principal reason one would have for thinking that the rational justifications for the cyclist's movements are not the cyclist's reasons is one's belief that those justifications do not figure in the cyclist's experience. That is, the analogy upholds the justifying role of experience, but insists that the justifications in question are insufficient, in that they are non-experiential. If the analogy were to be

[2]That something is an exculpation does not itself imply that it is not also a justification. Compare: that an object bears a causal relation to a subject does not imply that it does not also bear an intentional relation to that subject. If it were not true that we could see subject and object as bearing both kinds of relation toward each other, then the naturalist project would be doomed from the start. McDowell realizes that he must do more than show non-conceptual experience to be an exculpation; he must show that it is a *mere* exculpation. That is, he must show that it is not a justification.

correctly made to the case of empirical belief, then, McDowell would have to be attacking a view of the following sort: an empirical belief is justified by some extra-conceptual, non-experiential facts, such as facts about the belief's aetiology, the reliability of the perceptual mechanisms involved in forming the belief, etc. Then McDowell could point out that these facts do not figure in the experience of the subject that has the belief. So, as in the case of the cyclist, the extra-conceptual justifications do not constitute the subject's reasons.

But this is not the position being taken by a supporter of non-conceptual content (we do not have an analogy, but a disanalogy). For that position takes the belief to be justified by (non-conceptual) experience itself, not some extra-experiential facts that can be shown not to appear in a subject's experience. Since the justifications are (the contents of) experiences themselves, there is no way to deny their being the subject's reasons by claiming that they do not appear in experience. The analogy is of this form: in the case of the cyclist, it is the non-experiential nature of the justifications that shows that they are not the cyclist's reasons. *Since* the justifications of belief in terms of non-conceptual content are also non-experiential, we can infer analogously that they will also not constitute a subject's reasons for holding those beliefs. Analogies provide a major premise that, together with an independently established minor premise, motivate a conclusion. McDowell has not independently established that there is no experience involved in the case of non-conceptual justifications of belief. McDowell can only use the analogy as an argument if he *assumes* that the non-conceptual cannot be experiential.

McDowell takes the non-conceptual to be non-experiential. To put forward the analogy is to ask: why should we think that non-conceptual contents are the contents of *experience*, as opposed to being, e.g., sub-personal contents? McDowell is asking for a motivation: why should he not think that whatever possesses non-conceptual content is just as non-experiential as the cyclist's justifications, despite the stipulation to the contrary? His remarks (on p 164, and on p 166) suggest that one reason why he requires such a motivation is this: he does not see how a content can be canonically individuated non-conceptually. That is, he believes that any concepts used in canonically individuating a content must be possessed by the subject who entertains that content. And, more to the point, those concepts must be experientially active in any subject that has an experience with that content. This could be why he finds the idea of non-conceptual content so difficult to accept as the content of experience, since he believes (correctly) that the concepts employed, e.g., by Peacocke in individuating such contents are not active in the experiences these contents are intended to capture. Surely the concepts used in specifying the justifications for the cyclist's movements (*centre of gravity*, *desired trajectory*, or what have you) need not be possessed by every cyclist for which those justifications apply. So they cannot then be used to specify the cyclist's experience. But, as will be shown (see chapter 5), there are means of content individuation that are non-conceptual (e.g., Peacocke's scenarios): there is no requirement that the concepts employed in these specifications should be present in an experience that has that content. Once this is seen, the analogy of the cyclist can be dismissed as irrelevant.[3]

4.3 Non-experiential reasons

McDowell's insistence that all experience is conceptual is not trivialized by a weak notion of the conceptual. Conceptuality, for him, requires a capacity "to decide whether or not to judge that things are as one's experience represents them to be" (p 11). This requires a considerable degree of objectivity: a grasp of the subject/world distinction, the seems/is distinction, a grasp of the concept of truth. And further: "conceptual capacities... belong to a network of capacities for active thought... [which] takes place under a standing obligation to reflect about the credentials of the putatively rational linkages that govern it" (p 12). And again: it is essential to the capacity to have an experience with conceptual content, that it can be "exploited in active thinking, thinking that is open to reflection about its own rational credentials" (p 47). McDowell is arguing, then, that there is

[3] If McDowell were right in saying that the rational linkages he mentions are not the cyclist's reasons for making the movements, then it would seem that the cyclist has no reasons for those movements. Thus, without a notion of non-conceptual content which could admit those linkages as reasons, no personal level explanation of the cyclist's movements can be given. Yet it may be that many, if not most, modes of cognitive activity are at heart like riding a bicycle: skilled, real-time coordination with one's environment through integrated perception and action. This is another illustration of how generalizing our notion of content to include the non-conceptual can dramatically increase the range of psychological explanations we can provide.

no (experiential) content in an organism which lacks these distinctions, concepts, and reflective abilities. This is important, for if McDowell left unspecified just how rich a notion of conceptuality is in play when he denies experience to those organisms lacking a conceptual life, his position would be underspecified to the point that one might think there is no tension between his position and mine. But there is a difference: I believe that organisms without those sophisticated capacities may nevertheless have experience with content (or at least, that there are some experiences the having of which is not by virtue of possessing the conceptual capacities McDowell mentions). There are experiences the content of which is not conceptual.

In fact, one can see McDowell's principal mistake as insisting (rightly) on such an elevated notion of the conceptual, and yet maintaining that it is only within the conceptual sphere that our thoughts can find justification:

> We cannot really understand the relations in virtue of which a judgement is warranted except as relations within the space of concepts: relations such as implication and probabilification, which hold between potential exercises of conceptual capacities [McDowell, 1994b, p 7].

In the light of the elucidation of the notion of non-conceptual content in chapter 3, and the examples of phenomena that require a notion of content (defined, partly, in terms of implication relations) but do not support the ascription of objectual content (chapters 2, 7 and 8), McDowell's assumption, that the relation of implication can only hold between conceptual elements, begs the question.

However, the difference between devotees of non-conceptual content and McDowell may not be as stark as this suggests. For example, in denying "inner experience" to organisms lacking the conceptual sophistication described above, he says:

> I have been claiming that it is essential to conceptual capacities that they belong to spontaneity... Whatever it may be that is true of a creature without spontaneity when it feels pain, it cannot be true that it has "inner experience", according to the picture of experience that I have been recommending [McDowell, 1994b, pp 49-50].

He similarly rejects the possibility of "outer experience" in such creatures. But notice that he allows that they can "feel pain", despite their lack of conceptual sophistication. Surely such feelings would count as experience, under a non-Kantian understanding of "experience". McDowell finds a tight connection between conceptuality and experience, not by weakening his notion of the former, but by bolstering the latter substantially: "inner experience" might be absent even though there is a feeling of pain; "outer experience" might be absent although there is perception. Therefore, even if experience in McDowell's restricted sense must have only conceptual content, this leaves open the possibility that the content of experience in a more general sense might be non-conceptual.

The preceding section, in responding to McDowell, noted that even if one conceded an essential role for experience in providing reasons, it is still possible for non-conceptual contents to provide reasons, because experience can be non-conceptual. On that view, McDowell's account of the cyclist as not acting for a reason is correct, but irrelevant, since the justifications involved in that case are (supposedly) not experiential, yet the justifications offered by non-conceptual content *are* experiential.

A different response to McDowell's analogy would be to use his restricted notion of experience against him, and deny what he says about the cyclist. One could insist that the mentioned rational justifications may constitute the cyclist's reasons, even if they do not enter into experience, as commonly understood, at all (neither conceptually nor non-conceptually). Despite the fact that these justifications are non-experiential, they nevertheless rationalize the cyclist's actions. Certainly we, as persons, have many (content-involving) reasons which do not enter into our experience (implicit or supporting beliefs are one clear case). That is so even on the common, and relatively weak understanding of what experience is. A fortiori, then, there can be reasons which are not experiential on McDowell's more restricted understanding of what experience is.

Consider this example: although the belief *someone could get hurt by this* might not enter into my experience when I leave a hole in the pavement unmarked, I would still under normal circumstances be ascribed the belief and therefore be held responsible for any injury caused. In fact, I would be held culpable precisely because I had the belief, and yet it did not enter into my experience and guide my action. Contrast this with the leniency with which we would treat someone (say, a child) for whom we have grounds to think they do not hold the implicit belief *someone could get hurt by this*. The contrast shows that the content of the belief, when present, is not sub-personal, below the level of action. Rather, the content is personal content, yet non-experiential. So the analogy of the cyclist is rejected: experience is not required for personal content. So we can understand how non-conceptually individuated contents may be the contents of our personal life, even if the concepts in those individuations, or the contents themselves, do not enter into our experience. If the conceptual is the experiential, then the contentful does indeed stretch beyond the space of concepts.

But the point of McDowell's analogy is this: the contentful, in a loose sense, indeed extends beyond the space of concepts. Perhaps even personal-level content does. But that does not mean that reasons do. That there are non-experiential personal contents does not imply that there are *reasons* which are non-experiential. Reasons guide and rationalize action. The example of the hole in the pavement highlights the *failure* of the implicit belief to rationalize and guide my action. It focuses on a personal-level content that is exactly *not* a reason.

Nevertheless, the digging example can accommodate this. For I engaged in the activity of digging, no doubt, for a number of reasons, many of which were non-experiential. For example, it seems very likely that one of my reasons for digging was that I wanted to make a hole and I believed that *digging yields holes*. Yet it seems quite possible that there was no episode in my experiential life around the time of the digging that had the belief content *digging yields holes*. On a weak notion of experience, perhaps; on McDowell's bolstered notion, no.

But perhaps this is to misunderstand McDowell's (neo-Kantian) notion of experience. Perhaps, on his view, all that one *means* when one says that a belief of mine is experiential is that the belief provides a reason for my action. McDowell would, no doubt, claim that by experience he does not mean something that is primarily understood in terms of private, inner experiences, qualia, phenomenology, etc. So even if any of these are not concurrent with my belief *digging yields holes*, that in itself does not exclude that belief from my experience. So an example of non-experiential reasons has not been provided.

But then it is hard to imagine what *could* count as a counter-example. For with this understanding of experience, it is *stipulated* that all reasons are experiential. And there is a familiar price to pay for stipulation. Specifically, McDowell can then not appeal to some independent notion of experience that is not satisfied in the case of the cyclist. If we have prima facie evidence that there are reasons involved (which we do, by virtue of the presence of rational linkages), then the stipulation implies that we therefore have prima facie evidence that there is experience involved. With the stipulative connection between experience and reasons, one can only question the presence of experience by questioning the presence of reasons directly. The questioning of the presence of reasons cannot proceed via a prior questioning of the presence of experience, for that would defy the stipulation. It would suppose that one means something more than facts about reasons when speaks of experience. In short, without the stipulation, we have no reason to believe that reasons should only be present in experience, and the cyclist analogy fails. Yet with the stipulation we no longer have a reason to doubt that the cyclist's rational linkages are experiential.

Perhaps McDowell could admit this, but claim that it does not tell against his argument against non-conceptual content. In showing that there are non-experiential reasons, I appealed to beliefs such as *digging yields holes*. Yet surely such a belief is a conceptual one. Thus, I may have demonstrated that the space of reasons extends beyond the experiential, but I have not shown that it extends beyond the conceptual.

So one might think, but this would be to misunderstand the dialectical role of the digging example. Its role is to provide a counter-example to the general premise that all reasons must be experiential. Without that general premise, one cannot use the cycling analogy to argue against non-conceptual reasons and non-conceptual content.

If what I have said here is right, then McDowell's mistake is to locate the person in the (experiential) spontaneity. He makes this mistake because of his adherence to the other half of the Kantian slogan: "concepts without content are empty". (McDowell makes it clear (p 3-4) that what Kant meant by "content" here is more like what McDowell means by "experiential intake".) The idea is that thoughts that are not grounded in experience will fail to make referential contact with the world, will be about nothing. (McDowell admits that this is only a constraint for "empirical thinking", for other forms of thought are not meant to have such connection to the world.) Thus, there can be no reasons that are not experiential.

The alternative idea of non-experiential, yet personal, content is thus the idea that we do not need an experiential intermediary to make referential contact with the world. Justification from the world does not need to be channeled through an experiential bottleneck before reaching the thought it grounds. On this possibility, McDowell is silent.

4.4 Inarticulable reasons

We have seen (in his discussion of the analogy of the cyclist) that the kind of justifications that McDowell is after are reasons. We have also seen that McDowell insists on two restrictions on reasons: they must be conceptual, and they must be experiential. McDowell further restricts the notion of reason he has in mind when he makes another objection to non-conceptual contents as reasons:

> In the reflective tradition we belong to, there is a time-honoured connection between reason and discourse. We can trace it back at least as far as Plato: if we try to translate "reason" and "discourse" into Plato's Greek, we can find only one word, *logos*, for both. Now Peacocke cannot respect this connection. He has to sever the tie between reasons for which a subject thinks as she does and reasons she can give for thinking that way. Reasons that the subject can give, in so far as they are articulable, must be within the space of concepts [McDowell, 1994b, p 165].

In responding to this, as to many other claims, one can take a hard or soft line; and which line one takes will depend on the reading one gives to McDowell's term "articulation". The hard line is to deny outright that articulability implies conceptuality. One could see chapter 5 as taking this line. There I admit that there are some obstacles to articulating non-conceptual contents (e.g., such contents cannot be specified by conventional "that" clauses). But in that chapter I overcome those obstacles, and present several means of specifying vertical non-conceptual contents (as Peacocke has done for horizontal non-conceptual contents). The hard line, then, sees specification or individuation as sufficient for articulation.

The soft line grants McDowell a more restricted notion of articulability. On this understanding, mere individuation is not sufficient for articulation. Articulating a content is *expressing* it, uttering an expression which has the very same content as the one to be articulated. It is clear that this is the notion of articulability that McDowell has in mind:

> The idea [that the space of reasons is more extensive than the space of concepts] is that when we have exhausted all the available moves within the

> space of concepts, all the available moves from one conceptually organized item to another, there is still one more step we can take: namely, pointing to something that is simply received in experience. It can only be pointing, because, *ex hypothesi* this last move in a justification comes after we have exhausted the possibilities of tracing grounds from one conceptually organized, and so articulable, item to another [McDowell, 1994b, p 6].

Thus, McDowell allows that although one may not be able to articulate an extra-conceptual item, one can point to it. Yet this is all the advocate of non-conceptual content needs. The means of individuating non-conceptual contents used in chapter 3, and the means of specifying them in chapter 5, point to such contents, rather than articulating them. Those means do not *express* the content at hand, but *refer* to it. On this understanding of articulability, the defender of non-conceptual content can certainly accept that true articulability does indeed imply conceptuality (in fact, a premise of this sort is behind at least one argument (in chapter 5, section 5.2) that one cannot specify non-conceptual contents via "that" clauses). This can be accepted because articulation is not the only means of indicating a content.

Furthermore, even such an advocate of non-conceptual content need not deny the time-honoured "connection" between reasons and articulables. To acknowledge a connection between the two is not to equate them. For example, there is good reason to believe that an account of concepts following Peacocke's strategy would include, either implicitly or explicitly, articulability requirements in the possession conditions for many concepts. This alone could account for the presence of a connection between the two in our "reflective tradition". But it would not prohibit the possibility that some concepts could be possessed and applied for good reasons without the subject being able, even in principle, to articulate those reasons. No doubt McDowell has a stronger notion of "connection" in mind. But even if believing that all reasons are articulable has been the view of our tradition, he does not show that it is the correct view.

4.5 Non-conceptual content and modesty

So far in this chapter I have been on the defensive, responding to McDowell's objections to non-conceptual content, to show that it, like conceptual content, can provide grounding for our empirical thinking. But if conceptual content can already provide this grounding, then the possibility of non-conceptual content is less interesting. Why bother trying to make clear this strange notion of non-conceptual content when we already have a notion of content that does the job?

Of course, I have been on the offensive in other chapters, arguing that there are limitations to conceptual content, especially concerning psychological explanation, that non-conceptual content can transcend. But even in the areas of McDowell's concern, there is a limitation to conceptual content that makes the non-conceptual account more attractive. Specifically, any purely conceptual justification of our empirical thought must be circular.

McDowell admits that in giving an account of an observational concept, one will not be able to appeal to only *other* concepts. Rather, since the account must be conceptual, it must be that the account employs the very concept of which an account is being given. Furthermore, there is a "primitive fragment " of our conceptual repertoire such that an account of any concept in that fragment will be employed within the scope of the propositional attitudes of a subject, yielding a kind of circularity: in giving an account of such a concept, it is assumed that one already knows what it is for someone to take attitudes toward that concept. That is to say McDowell thinks that our accounts of concepts must be *modest* [McDowell, 1987]: we cannot give a non-circular account of all concepts. McDowell therefore correctly locates one attraction of a notion of non-conceptual content as deriving from the desire for a non-circular justification of thought:

> What a good account [of the concept F] must avoid is ineliminable mention of the concept F as the concept F within the scope of the propositional attitudes of the thinker. If the account does mention the concept in *that* way, it will not have elucidated what it sets out to elucidate [Peacocke, 1992, p 9].

This desire for non-circularity naturally suggests a notion of non-conceptual content:

> When we set out to give an account of what it is to possess, for instance, the concept *red*, we shall find ourselves saying things like this: to possess the concept *red* one must be disposed... to make judgements in whose content that concept is applied predicatively to an object presented to one in visual experience, *when the object looks red to one*, and for that reason [McDowell, 1994b, p 167].

If this notion of an object looking red to one can only be understood as conceptual, then the non-circularity requirement will not have been maintained. One can only avoid circularity if one is able to understand, for example, an object's looking red to someone as being a justification, yet not involving the concept red. One can only avoid circularity if one has the notion of a non-conceptual justification. We have seen how McDowell tries (if I am right, unsuccessfully) to show that such justifications are impossible: justifications must be reasons, and reasons must be conceptual.

But what exactly is McDowell's modest alternative? A modest account is not just one which, in giving an account of a concept, *uses* that concept: Peacocke's favoured accounts do this. The point of disagreement is whether or not the concept can be mentioned "within the scope of the propositional attitudes of the thinker". Modesty is the claim that, for at least some concepts, there is no account of them does not use those concepts within the scope of the subject's propositional attitudes. By why should that lead to any objectionable circularity? The objectionable aspect must be this: a modest account of a concept requires that the theorist already have the ability to ascribe that concept to others.[4]

To say that a modest account is circular is not to say that it is uninformative. An account of the concept *red* that employs the notion of a subject judging something to be red because it looks red, although circular, *does* tell us something. Evans distinguished between three elements of the informational system: perception, memory, and testimony [Evans, 1982, p 122]. A thought may be informationally grounded through any one of these modes, or through a mixture of modes. Therefore, the modest account is informative in that it distinguishes the actual grounds for the judgement (perceptual) from other possible grounds: recognizing the apple as one previously experienced (memory); or having been told that the apple is red (testimony). But this is not enough to uniquely identify, in a non-question-begging way, what is characteristic of the concept *red*.

Peacocke and McDowell agree, that a modest account of concepts is circular, at least for observational concepts; see [McDowell, 1994b, p 169]. However, on an understanding of modesty as requiring an ability to ascribe the concept being elucidated, there are some

[4] Or, in the case of indexical concepts, the circularity of modesty is, in giving an account of a concept type (such as the first person), the presumption of a prior ability, on the part of the theorist, to ascribe instances of that type to others.

doubts that modesty really is circular.

First, there is the thought that modesty will only yield circularity if one assumes Peacocke's Principle of Dependence: there can be nothing more to the nature of a concept than an account of what it is to possess that concept [Peacocke, 1993, p 5]. For if that Principle were not assumed, one could believe that there are accounts of the nature of concepts that do not attempt to give an account of "*what it is* to possess a concept". Such accounts could assume an ability to ascribe the contents (knowledge of what it is to possess a concept), and yet elucidate whatever it is about the nature of concepts that supposedly transcends their conditions of possession.[5]

Furthermore, it seems possible for one correctly to ascribe a concept to others, without knowing *what it is* to possess that concept (surely this is what we do every day). If so, then even assuming Peacocke's Principle of Dependence, a modest account would not necessarily be circular in any disadvantageous sense. For in such a case knowing *what it is* to possess a concept would not presuppose that very knowledge, but rather an ability to ascribe the concept reliably (because modest accounts employ the concept within the scope of a subject's propositional attitudes), which can be had (obviously) without the deeper understanding of the concept's nature.

But even if a modest (purely conceptual) account of concepts is not *ruled out* by its circularity, there are still reasons to prefer an ambitious (non-conceptual) account over a modest one. In not presupposing a prior ability to ascribe the concept to be elucidated, an ambitious account may attempt to *provide* such an ability *by virtue of* a theoretical knowledge of what it is to possess the concept. This would be of use for some of the types of psychological explanation with which this thesis has been concerned, such as explaining

[5] According to this understanding of modesty, the conceptual subtraction method of content specification (see chapter 5, 5.3.1) would be a modest one. That is, if contents are to be specified by indicating a conceptual content and the differences between that conceptual content and the content being specified, then it seems unavoidable that there will be some conceptual contents whose specification by this method will be pleonastic, and thus circular. However, employing a modest means of specification does not require one to employ a modest means of individuation; one may use the conceptual subtraction method and still have a non-circular account of content. For the difference between individuation and specification, see chapter 5, section 5.2.

the behaviour of those who have concepts which we do not have (or with which we lack practical familiarity).

McDowell explicitly states his dislike of such accounts. In that they are not modest, they are "from sideways on":

> What I do mean to rule out is this idea: that, when we work at making someone else intelligible, we exploit relations we can already discern between the world and something already in view as a system of concepts within which the other person thinks; so that as we come to fathom the content of the initially opaque conceptual capacities that are operative within the system, we are filling in the detail in a sideways-on picture – here the conceptual system, there the world – that has been available all along, though at first only in outline. It must be an illusion to suppose that this fits the work of interpretation we need in order to come to understand some people, or that a version of it fits the way we acquire a capacity to understand other speakers of our own language in ordinary upbringing [McDowell, 1994b, p 35].

I must interrupt McDowell here and say that, although what he has just said may be true, it does not tell against the prospects of a sideways-on picture of content for the purposes of psychological explanation of behaviour. Even if we do not use the sideways-on picture in our everyday efforts in understanding others, or even in our childhood efforts in language learning, it still may be the best picture to use in our efforts to understand each other in a more objective, scientific way. We should not assume that the modes of interpretation that have had the most pragmatic value for non-theoretical purposes are the ones that reveal content and concepts for what they really are.

McDowell continues his criticism of the sideways-on picture:

> The illusion is insidious; so much so that it can entice us into aspiring to a sideways-on understanding of our own thinking, which we take to be the condition of someone else who understands us. *Some* sideways-on picture must be innocuous in the case of a thinker who is opaque to us, and then it can seem obvious that overcoming opaqueness is just filling in blanks in that sideways-on picture, leaving its orientation unchanged. But that must be wrong. The mistake is not to give proper weight to this fact: in the innocuous sideways-on picture, the person we do not yet understand figures as a thinker only in the most abstract and indeterminate way. When the specific character of her thinking starts to come into view for us, we are not filling in blanks in a pre-existing sideways-on picture of how her thought bears on the world, but coming to share with her a standpoint *within* a system of concepts, a standpoint from which we can join her in directing a shared attention to the world, without needing to break out through a boundary that encloses the system of concepts [McDowell, 1994b, pp 35-36].

McDowell's rejection of a sideways-on view does contain a truth, just not the one McDowell thinks it does. We can adopt the external view with respect to the concepts of others, but only to a certain extent. The limitation of that view is determined by the realm of reference. It is with respect to reference that we cannot take the sideways-on view. We must assume that we share a common world with the subject whom we wish to understand. Even if we possess different concepts, different ways of experiencing that world, our different concepts nevertheless make contact with the very same objects and properties.

The demand for a rejection of a sideways-on approach to reference is made vivid in considering non-conceptual content. If the Generality Constraint is characteristic of concepts, then to make sense of non-conceptual content we must make sense of constituents of content for which that Constraint does not hold. But a familiar point about the Constraint is that it only seems true if one assumes some prior set of referential (or "semantic") categories, in the context of which the Constraint may be defined. Otherwise, the Constraint would be false for all constituents, even those which we intuitively wish to call concepts, since for any mode of presentation of a particular a there is always a mode of presentation G of a property, such that $G(a)$ is semantically anomalous (e.g. *the building is prime*). By understanding the subject as inhabiting the same referential realm we do, we are free – no, we are compelled – to use our own referential categories as the determiners of the conceptuality of their content constituents (see chapter 3, section 3.3.4).

The sideways-on view of reference is not one that can be maintained, and it is this fact that gives the above extended passage from McDowell the ring of truth that it has. But that is not to say that a sideways-on view of content is similarly untenable; on the contrary, it is essential to understanding others different from ourselves.[6]

[6] I do not mean to suggest that every ambitious (non-modest) account must reject what McDowell is calling the sideways-on view. Peacocke would no doubt maintain that his non-modest possession-conditional account can take place from within, not without, the system of concepts of the thinker we are attempting to understand.

4.6 Sub-personal vs. sub-organismal content

McDowell admits that he has only one particular notion of non-conceptual content as his target:

> I am not saying there is something wrong with just any notion of non-conceptual content. It would be dangerous to deny, from a philosophical armchair, that cognitive psychology is an intellectually respectable discipline, at least so long as it stays within its proper bounds. And it is hard to see how cognitive psychology could get along without attributing content to internal states and occurrences in a way that is not constrained by the conceptual capacities, if any, of the creatures whose lives it tries to make intelligible. But it is a recipe for trouble if we blur the distinction between the respectable theoretical role that non-conceptual content has in cognitive psychology, on the one hand, and, on the other, the notion of content that belongs with the capacities exercised in active self-conscious thinking – as if the contentfulness of our thoughts and conscious experiences could be understood as a welling-up to the surface of some content that a good psychological theory would attribute to the goings-on in our cognitive machinery (p 55).

Thus, McDowell only takes issue with anyone who is proposing that non-conceptual content can be involved in active, self-conscious thought; other uses of the the notion of non-conceptual content are unobjectionable. It would thus seem that given my interest in applying non-conceptual content to pre-objective intentionality, I could welcome this passage from McDowell as exempting from his criticism my notion and use of non-conceptual content. But someone with my concerns should not accept the passage, for several reasons.

First, the offered compromise – the neat division of conceptual content for active self-conscious thinking, non-conceptual content for lesser states – should be rejected. Although my argument for the need for non-conceptual content has, for simplicity, granted that conceptual content and active, self-conscious thought go hand-in-hand, it remains possible that some active, self-conscious thought may have non-conceptual content. Furthermore: although no one, so far, has successfully argued for the non-existence of conceptual content, it nevertheless should be acknowledged that conceptual, not intentional, eliminativism is an open possibility that should be investigated. Accepting McDowell's compromise would be to ignore this issue.

Even if one were to restrict non-conceptual content to thought that is less than active, self-conscious thought, this would not mean that it must only play a role in sub-personal explanations, at least in the sense of "sub-personal" that McDowell has in mind. For him, content at the sub-personal level is "irreducibly metaphorical" [McDowell, 1994a, p 197]; it is the content of mechanisms, states and processes that are not subjects of experience. I agree with McDowell that non-conceptual content will be of essential use in such explanations, since typically they will demand a type of content for which conceptual constraints (such as the Generality Constraint) do not hold. But I have argued that there is a type of psychological explanation which neither presumes a reflective, self-conscious, conceptual subject, nor does it involve only sub-personal states. Rather, such explanations concern the contentful states of an experiencing organism, albeit one that is lacking significant conceptual abilities (or is not exploiting them in the mental episode that is being explained).

This point can be put another way. I said that the above remarks assume McDowell's sense of the term "sub-personal". The fact is that I have recently come to use that term differently than is the current custom. I might, on my new understanding of the term, agree that all thought below active, self-conscious thinking is indeed sub-personal. But I would only consent to that if another distinction is acknowledged, the organismal/sub-organismal (or organismal/organal) distinction. On this new understanding, an experience may be sub-personal, in that it is the experience of an organism that is less than a person, less than a self-conscious subject with full conceptual capacities. Yet the content of that experience is not the content of mechanisms, processes and internal states (which would be sub-organismal, or organal, content); it is the content of an organism's experiences, so it is organismal content. In the terms of this new nomenclature, McDowell assumes that all sub-personal content is sub-organismal. But it is not.[7]

Last, I object to McDowell's compromise since it unduly denigrates the role of sub-personal (in my terms: sub-organismal) explanations with respect to personal- (organismal-) level ones. I agree with McDowell that the Dennettian picture of organismal contents just being sub-organismal contents that "well up" is an unattractive one (though I am not convinced that it is refuted in [McDowell, 1994a]). However, many more complex relationships between the two levels are possible, and have yet to be explored. It is some such

[7] See [Elton, 1995] for another example of someone who has found the organismal/organal distinction useful.

story, not the “welling-up” straw man, that may allow us to naturalize the experiential in terms of the non-experiential, via the intermediary notion of non-conceptual content.

CHAPTER 5
Practical specifications of content

5.1 Introduction

This chapter claims that the usual "that"-clause specification of content will not work for non-conceptual contents; some other means of specification is required, means that make use of the fact that contents are aspects of embodied and embedded systems. In particular, the development and deployment of the notion of non-conceptual content requires assistance from a practical and theoretical understanding of computational/robotic systems acting in real-time and real-space.

5.2 The inadequacy of standard specifications

In order for a theory of intentional action to be able to appeal to specific contents in its explanations, it must have a means of canonically specifying those contents, a means of specifying them according to their essential properties, such as their semantic values or their psychological significance (viz. Cussins: "Something is canonically characterized (within a theory) if, and only if, it is characterized in terms of the properties which the theory takes to be essential to it" [Cussins, 1990, p 382, fn]). For example, one can specify a content by the phrase "the content toward which subject A took the belief attitude exactly 10.3 seconds ago", but this would not be a canonical specification, since it does not pick out the content it does in virtue of the content's essential properties, but rather (some of) its accidental ones.[1]

[1] I am using the term "canonical" in a slightly idiosyncratic way, then, since on my view it is possible for there to be more than one canonical means of specification of a content (although there will be, in

Chapter 3 argues that conceptual content is individuated with respect to connections to judgement, and non-conceptual content is individuated with respect to those connections plus others, such as developmental relations between contents, and a small example is provided for each (sections 3.2 and 3.7, respectively). However, the extremely theoretical nature of this individuation scheme renders it fundamentally inappropriate for the practical explanation of particular intentional phenomena. Consider how infeasible it would be to specify conceptual contents in our explanations of each other by enumerating their logical relations to other contents, their acceptance conditions, etc., rather than by using standard "that"-clause specifications. Likewise, a practical, yet canonical means of specifying non-conceptual contents is required.

5.2.1 Linguistic use specifications

A standard means of canonically specifying contents is what I will call the "linguistic use" means of specification: providing an expression in English (or other natural language), usually preceded by "that", with the same content as the one to be specified (e.g., "The content of my belief is that the object on the table is a computer" or "The last bit of register A0 being on means that a message is in the input buffer").

Of course, almost any proposed means of specifying content will use language in a more general, and conventional, sense of the word "use" than the one I am employing here. However the following criticisms of linguistic use methods are not directed towards proposed forms of specification that use language in this broad sense. The expression "linguistic use" is meant to be a technical one: I mean to include under the term only those means of specification that pick out contents exclusively in the manner mentioned in the definition above: viz., by providing an expression in a natural language that has the very same content as the one that is to be specified. In this narrower, more technical sense, means of specification may use language in the general sense, and yet not be subject to the negative conclusions in what immediately follows.

Although the linguistic use means of specification might work well for conceptual contents, there are several reasons, given in the rest of this section and in the next, why

general, only one specification of the content per canonical means).

one might think that it is not adequate for the specification of non-conceptual contents (NCCs).

First, some might wish to dismiss arguments to the contrary [Travis, 1994], and claim that language is itself conceptual, and therefore linguistic use specifications can only specify conceptual contents. Linguistic use specifications are what have been called elsewhere (e.g. [Cussins, 1990, p 382 ff], [Peacocke, 1986, p 17] and [Crane, 1992, p 142]) conceptual specifications of content: specifications that are made in such a way as to require a subject to possess the concepts used in the specification if that subject is to be able to take an attitude toward that content (see chapter 3, section 3.3.1). Thus linguistic use specifications, employing conceptual language, will not be able to specify the contents of, say, the infant from the Example (see chapter 2), since such specifications would require the infant to possess concepts which it in fact lacks. This is because in order to entertain a content that has C as a conceptual constituent, one must possess the concept C (see chapter 3).

In spite of the strong intuitions behind this line of thought, there are reasons why it might be more illuminating to establish the incompleteness of linguistic use specifications by a means other than one which relies on the principle that all language is conceptual; for one thing, some would want to deny that language is entirely conceptual [Travis, 1994]. What I do, then, is split linguistic use specifications into two types: purely descriptive, and indexical. I argue that specifications of neither type can specify NCCs. First, a means of specification, in order for it to be of use in a scientific theory, must specify NCCs canonically, which rules out descriptive linguistic use. Furthermore, content specifications must be context-independent, which rules out indexical linguistic use. Thus some means of specification other than linguistic use is required.[2]

[2] Peacocke [Peacocke, 1990] shows the insufficiency of standard specifications for a restricted class of NCCs, perceptual demonstrative contents; and develops (in [Peacocke, 1989] and [Peacocke, 1993, ch 3]) an alternative means of specification, scenarios, for these contents. But the goal in this chapter is to establish the insufficiency of linguistic use specifications for NCCs in general (or at least for a broader or distinct class of NCCs than does Peacocke); for these NCCs the scenario means of specification will not work (nor was it intended to).

5.2.2 The inadequacy of descriptive linguistic Use

Cussins [Cussins, 1990] has brought together some insights from Evans [Evans, 1982] and Perry [Perry, 1979] that can serve as an argument against the possibility of using descriptive (i.e., non-indexical) language to specify non-conceptual contents.

Perry shows that there are contents, constitutively linked to perception and action[3] (e.g. the contents of one's "I" thoughts), that are not equivalent in terms of cognitive significance to any contents specified, in a purely descriptive manner, by the linguistic use method.

One can give a linguistic use specification of the contents of one's "I" thoughts, but only if one employs indexicals, as in the ascription: "RC believes: 'I am spilling sugar all over the supermarket floor'". One cannot use a non-indexical specification, as in "RC believes 'the person named RC is spilling sugar all over the supermarket floor'", since it specifies a content that is distinct from that of the first-person thought I would normally have in that situation, as can be seen by the differences in the two contents' connections to action (due to amnesia, I might think in the latter case "Well, the person named RC had better clean it up" and go on my way, whereas in the former case no amnesia could get me to think that it was anyone else's mess but mine). In order for the belief *the person named RC is spilling sugar all over the supermarket floor* to have any implications for my action, it must be supplemented by the belief *the person named RC = I (me).* The belief *I am spilling sugar all over the supermarket floor* requires no such further identification; its connections to action are direct, un-mediated.

The application of Perry's insight to the case of NCC is direct: if any NCCs are directly connected to perception and action in the way that "I" contents are, then Perry's arguments establish that such contents cannot be specified by means of descriptive language use. One line of reasoning that leads one to conclude that all NCCs are constitutively connected to perception and action is the following. As observed before (in chapter 3), NCCs are sub-objective in virtue of the fact that they do not enable the bit of the world

[3] Actually, Perry claims that it is the fact that a belief is a "locating belief" that makes its specification essentially indexical; I'm favouring here Cussins' analysis that it is a content's constitutive links to perception and action that requires non-descriptive specification for that content.

being thought about to be integrated into a unified framework of particulars and their inter-relations. It is this lack of a framework which restricts sub-objective thought to contents that are essentially linked to perception and action. The idea employed here is that indexicality is the starting point; contents that (merely) have constitutive connections to action and perception are the basic case. It is only through the construction of a non-solipsistic conception of the world via some unified framework of particulars and relations that one's contents can display the kind of perception and action transcendence that is characteristic of descriptive modes of thinking. It is the very sub-objectivity of NCCs that allows the application of Perry's argument to their case: they cannot be specified by descriptive language use.

Another way of seeing why it is that NCCs are constitutively connected to perception and action is to recall from chapter 3 that NCCs are individuated not just with respect to cognitive significance, but with respect to psychological significance in general. Psychological significance includes all normative transitions to and from contents; this makes direct links to perception and action at least partly constitutive of non-conceptual contents in general.

This understanding of NCC seems to agree with (at least one reading of) what Evans meant by non-conceptual content:

> Let us begin by considering the spatial element in the non-conceptual content of perceptual information. What is involved in a subject's hearing a sound as coming from such-and-such a position in space?... When we hear a sound as coming from a certain direction, we do not have to think or calculate which way to turn our heads (say) in order to look for the source of the sound. If we did have to do so, then it ought to be possible for two people to hear a sound as coming from the same direction (as 'having the same position in the auditory field'), and yet to be disposed to do quite different things in reacting to the sound, because of differences in their calculations. Since this does not appear to make sense, we must say that having spatially significant information consists at least partly in being disposed to do various things [Evans, 1982, pp 154-155].

This is very similar to Perry's way of characterising ways of thinking that are essentially linked to perception and action. Just as there is no "calculation" in the case of Evans' example of auditory content, there is no "calculation" in Perry's example of the first-person mode of thought: one knows, in an un-mediated manner, that such thoughts are directly related to one's own actions and perceptions. An identification with some descriptive mode of thought (e.g. *I (me) = the person named RC*) is not required for action.[4] Because of this, linguistic expression with entirely descriptive content can specify non-conceptual contents.

Another indication that NCCs are, like the contents of the indexicals "I" and "now", constitutively linked to perception and action is that if one attempts to specify such contents by means of linguistic use, then one tends to use indexicals in so doing.[5]

If what has been said is correct, then NCCs are indexical contents in Perry's sense, and therefore, like the content of conceptual first-person thoughts, cannot be specified by descriptive linguistic use. But there is reason to believe that non-conceptual contents, unlike the contents of conceptual first-person thoughts, cannot be specified by the alternative of indexical linguistic use, either.

This is so, if not because of the conceptuality of language, as discussed above, then for reasons related to the requirement that scientific theorising be context-independent (in a particular sense). That is, even if indexicals could, *per impossibile*, be used to specify NCCs via linguistic use (either because indexicals do not have conceptual content, or because they can somehow linguistically specify a content that is not, strictly speaking,

[4] A word of caution: Evans' point should not be construed to be claiming that non-conceptual contents are somehow infallible, because of their direct connections to perception and action. The essential links can be inappropriate for the current situation, therefore yielding a false NCC: because of a reflection, the sound might be heard as coming from the right (with all the commensurate right-directed dispositions), when in fact the source of the sound is straight ahead.

[5] Another caution: though I am arguing that all NCC's are indexical, in that they are non-descriptively linked to perception and action, I am not claiming that the relation in the other direction is true (that all indexical contents are non-conceptual, or sub-objective); on the contrary, I think "I" has conceptual content (I might be wrong on this, as on anything else, but fortunately it would have no undesirable consequences to what I am arguing here if I were). Unlike Cussins [Cussins, 1990, p 391, n 46], I do not feel justified in rejecting out of hand indexical linguistic use specifications of contents for a scientific psychology. Some indexical contents (the first-person, the present-tense) seem to be conceptual (at least enough to avoid the problems of context-dependence), and are thereby specifiable by indexical linguistic use. Thus, I am required to provide an argument (which I do) for the claim that indexical linguistic use specifications will not work for the case of non-conceptual contents (although embedded indexical specifications might succeed; see section 5.3.2).

the content they carry, or because it is possible to devise new indexicals that introduce, in a non-systematic way, elements of the environment into the content being specified) such indexicals alone would be inadequate for the particular task at hand: a context-independent intentional science. Specifically, the function of the indexical is merely to call attention to other factors (subject, context, and their relation) so that a content may be specified. In such a case, all the individuative work is being done by those highlighted elements, not the indexical itself.[6] Thus, in order to specify the content, one would need more than mere linguistic expressions; one would also need an environment related to those expressions in order to allow those expressions to function, and thus carry content. Thus, even indexical linguistic use cannot be used to specify NCCs.

These arguments agree with the conclusion of other writers (e.g., [Peacocke, 1981, p 191]): in specifying contents that are constitutively linked to perception and action, such as particular first-person modes of presentation, we cannot employ the content in question, but must refer to it instead. The task, then, is to find ways of referring to such modes of presentation that identify them not only uniquely, but canonically, as discussed above.[7]

[6] One might wonder: how is it that indexical linguistic use specifications seem to work in some cases, even though no systematic way of specifying various aspects of the context is at hand? The reply: in the case of indexical linguistic use specifications of conceptual contents (such as those that specify the mature first-person mode of presentation with "I"), the systematic, context-independent nature of conceptual thought permits specification with only one extra contextual parameter: the person grasping the content in the case of the first-person, and the time of the grasping, in the case of the present tense.

[7] Of course, one should evaluate other proposed alternatives to the standard means of linguistic use, few though they may be, before concluding that another alternative is needed, but there is no space here for such a survey. Suffice it to say that other alternatives (possible worlds, possession conditions, proto-propositional specifications) are insufficient either because they ignore cognitive significance (are purely extensional), they are impractical and unwieldy (as explained at the beginning of this chapter), or they are only faithful to the case of cognitive significance in the same cases (i.e., conceptual, or non-conceptual with shared environment) that linguistic use is, which I have already argued is insufficient for a scientific psychology.

5.3 Alternatives to standard means of specification

The rest of this chapter describes some possible alternative means of content specification that are currently under consideration, and explains why they are at least plausible candidates for an alternative. This not only serves to explicate many of the issues that have been discussed, and to pacify any "what else could there be?"-type worries; it will also be seen along the way why one must take embodiment seriously if one is to be able to specify non-conceptual contents for an intentional science.

5.3.1 Conceptual subtraction

One idea is: perhaps linguistic use fails only as a matter of technicality; perhaps some modification of it can overcome its limitations, while using the same, fundamentally non-embodied, approach. It seems such a modification would have to be something like the conceptual subtraction (CS) means of content specification. As the name might suggest, this method is similar in spirit to a pure conceptual specification, such as linguistic use. Nevertheless, and the above arguments against linguistic use specifications of NCCs notwithstanding, it seems possible that the CS method can specify non-conceptual contents, because it is distinct from pure conceptual specification in a crucial way. The CS method is an attempt to stay as close to our practice of linguistic use specification while throwing out the restrictions that make linguistic use inadequate. The problem with such purely conceptual specifications, as we have seen, is that they cannot specify non-conceptual contents. Any attempts at specifying the content of, say, the pre-objectual experiences of the infant from the Example, would over-ascribe, in that such specifications would imply that the infant possesses abilities that it does not, in fact, possess. That is, they would violate the Possession Principle (see chapter 3, section 3.2). For example, specifying the content of the infant's belief as "that there is a glass in front of me" would imply that the infant possessed the concept of a glass, with its attendant concepts, not only *drinking*, *manufacture*, and *glass*, but also *object* and *location*; it would also imply that the infant's thinking about the glass adhered to the Generality Constraint (see chapter 3), and supported the ability to think of the glass as something that could exist unperceived.

This would invite the theorist to make false predictions about, and would disallow correct explanations of, an infant's behaviour.[8]

The idea behind CS specification is to proceed with a conceptual, linguistic use specification, but also to tag the implications of that specification to which one does not wish to be committed; that is, to start with the conceptual content, and then subtract out properties of the conceptual content (such as "meets the Generality Constraint" or "supports the idea of existence unperceived") which the content to be ascribed does not possess. Then there will be no over-ascription of abilities, and therefore no false prediction or inaccessible explanation.[9]

The manner in which contents may fail to be objective is richly textured. A content may manifest its non-conceptuality by failing to respect any of a number of conceptuality constraints. In order to be able to employ the CS method, one must first capture all the different implications an ascription of a conceptual content carries with it. This will involve both a cataloguing of the general requirements for all conceptual contents and all concepts, such as the Generality Constraint, and a listing of the particular requirements for each individual concept.

First, a non-exhaustive and non-exclusive set of the former constraints might include, in addition to the Generality Constraint (GC), the following properties of conceptual content:[10]

(SP) it must have subject-predicate structure;

(RP) if the content is to be a way of thinking about an item, the subject must know

[8] This application of the Possession Principle assumes the Burgean qualifications made in chapter 3, section 3.3.2. That is, I am assuming that there is no reason to believe that the infant is able to exploit, through deference, any social contingencies that would warrant the ascription of a content whose concepts the infant did not possess. That is, I am assuming that the infant cannot supplement a partial mastery of a concept with a deference to a community in order to be capable of taking attitudes toward contents containing that concept. Therefore the Possession Principle applies.

[9] The basic idea of the CS means of specification seems to have been independently reached by Colin Allen [Allen, 1992].

[10] An example of a list of this kind, with several of the same entries, can be found in [Cussins, 1986, pp 218-219].

which item is being thought about (Russell's Principle; see section 3.6 of chapter 3);

(MC) if the content is to be a way of thinking about an item, the subject must be able to think of the same item in a number of other ways;

(EU) if the content is to be a way of thinking about an item, the subject must be able to think of the item as existing unperceived, as something for which the qualitative/numeric distinction applies, or as something which can be re-identified (see chapter 2, sections 2.4.3 and 2.4.4, and chapter 3, section 3.6) ;

(PE) if a subject is capable of taking the attitude of belief toward the content, then it must be able to entertain the possibility that its belief is false (or, slightly more plausibly, the possibility that this content might be false).

Next, a list of particular conceptual requirements might take the form of:

To possess the concept *bachelor*:
(1) subject must also possess the concept unmarried
(2) subject must also posses the concept male
...
To possess the concept *drinking glass*:
(1) subject must also possess the concept liquid
(2) subject must also posses the concept drinking
...
Etc.

Thus, non-conceptuality could be manifested in the failing to meet of any of these conceptuality constraints. It should be emphasised that these constraints are offered as examples only. The above lists are only meant to serve as a toy example of the enumeration of constraints for the CS method, and not as a specific proposal for what these constraints should be. It might be that some of the ones I happen to have included above are actually not required of all conceptual contents; or, it might be that some of them are required of all contents, and thus failing to meet them is not a way for a content to be non-conceptual. Also, the capturing of the commitments need not proceed via enumeration; the catalogue will undoubtedly employ quantification. It might also be recursive, in that conceptual requirements might themselves have further conceptual requirements, such that a content might meet some of the requirements for, say, (RP), and not others. Specifically, for the qualification that a constituent does not meet the Generality Constraint to be of any use, it would have to be supplemented with a specification of *how* the constituent fails to meet the Constraint. In chapter 3, section 3.4, it was explained how a constituent's failing to meet the Generality Constraint can be understood in terms of a gap between the constituent's range of application and its range of predication. Accordingly, the CS method can qualify how a constituent fails to be General by specifying this gap. Such specification need not be done purely extensionally, by listing each constituent that is in the range of application but not the range of predication. Rather, the specification of the gap may proceed quantifying over constituents. For example, we may define the application/predication gap g_1 to be: all subject constituents that are entertainable when the referent of the constituent is not perceptually occurrent.

Once this enumeration of conceptual requirements is in place, specifications of sub-conceptual contents in ascriptions would be possible:

> The content of the infant's belief is *that there is a drinking glass* $[-EU, -1]$ *within reach* $[-GC\{g_1\}]$

where the qualifiers within brackets after a concept indicate in what ways that part of the content fails to be conceptual, and the qualifier in braces after a -GC flag indicates the gap that characterizes how the Generality Constraint is not met.

In order for any alternative specification to succeed, several conditions must be met. Any particular application of the method must indicate at least one content, at most one content, and, as discussed before, it must indicate the content canonically.

In the case of the CS method, the first condition prompts one to wonder: how can one be sure that a subtracted content is actually a content at all? One suggestion[11] for a criterion for an abstract entity to be a content is that it be able to help rationalise a subject's behaviour by serving as a premise in practical reasoning. Thus, the need to meet this first condition highlights the fact that the CS approach only has meaning within the context of inference rules that relate such subtracted contents. A "logic" of subtracted contents is required, one that will capture the a priori relations between, e.g., the content "$glass[-EU, -2]$ at $location_1[-EU, -MC]$" entertained at time t_1 and the

[11] Thanks to David Charles for this suggestion.

content "$glass[-RP, -2]$ at $location_2[-EU]$" entertained at time t_2, where $location_1$ and $location_2$ are ego-centric specifications of places, such that they are co-referential, given the turning action performed between times t_1 and t_2. The inference from the first to the second content, in as much as it is correct, will have to fall under some inference rule in this "logic" of subtracted contents.

The second condition puts further constraints on the CS method. For it seems possible that there may be any number of ways that a concept could fail to incur some particular conceptual commitment. For example, it seems that any number of contents meet the condition "just like the concept *within reach*, but does not meet the Generality Constraint" (even with the gap qualifier), or "just like the concept *glass*, but does not support the ability to conceive of existence unperceived". So it seems that one's catalogue of conceptual commitments is going to have to be sophisticated indeed if one is to be able to specify a content uniquely.

But perhaps this just shows that the second condition is, strictly speaking, too strict. Of course, there is something to the idea that content specifications are useful only when there is some restriction on the contents that they specify. But this need not imply that specifications are of use in psychological explanation only when a unique content is specified. One might be able to specify only some restricted set of contents, those that share a particular set P of properties, as opposed to specifying a unique content. But if the explanation to be given need only appeal to the fact that the content possesses the properties in P, and if it can be made intelligible that the non-intentional characterisation of the system to be explained could instantiate some content with the properties in P, then perhaps no further individuation is required. In fact, anyone who thinks that many of our ascriptions of conceptual content are, strictly speaking, inaccurate will have to appeal to some consideration such as this in order to make sense of the fact that such ascriptions are as successful as they are.

With respect to the third condition, the CS method of specification will inherit the advantages of purely conceptual specifications: the ability to specify content in terms of its essential properties. If the first two conditions can be met or dispensed with, it seems that one can only question the canonicality of CS specifications if one is willing to question the canonicality of linguistic use specifications as well.

Another advantage of this close relation to conceptual specification is the ability to unify the conceptual and non-conceptual aspects of content within the same formalism.

But there are several obstacles to the successful deployment of this method. One possible worry is that the commitments to be subtracted must be atomistic: it must be the case that if one subtracts a commitment, one is not logically forced to subtract out other commitments. Or at least if there are such holistic inter-relations, they should be explicitly captured in a syntax of some kind. For example, if it is impossible to fail to meet the Generality Constraint without also failing to meet Russell's Principle, then either these should be rejected as candidates for commitments to be subtracted, or one must rule out, formally, the possibility of $C[-GC]$ (and $C[-RP]$) for all concepts C.

This worry seems unfounded, however. As long as the commitments P referred to in a specification are sufficient to meet the three conditions above, it doesn't seem necessary to refer to other commitments, even if they are holistically related to those in P. This view might have to be abandoned once one starts to develop a logic for subtracted contents, since one might want to guarantee, e.g., that distinct specifications imply distinct contents. But note that this is not guaranteed even for linguistic use.

But there are other worries. Perhaps there is no canonical, finitely-specifiable list of conceptual requirements, either in general, or for particular contents. Another possible difficulty is that the method might not be general enough; there might be non-conceptual contents that are not expressible as subtractions of conditions from conceptual ones. The problem is not that there might be non-conceptual contents that are subtractions of concepts other than those which we, as human theorists, possess; the fact that we do not possess these conceptual contents is not in itself an argument against the idea that they could be specified as logical functions of the concepts which we do in fact possess. Rather, the worry is that there might be non-conceptual contents that are not subtractions of any conceptual contents, be they in our possession or not. Without an argument against such a possibility, it would be excessively teleological to assume that all non-conceptual contents must be able to be expressed as subtractions from the conceptual contents into which some of them develop.

Finally, there is a general problem for non-embedded means of content specification, including both linguistic use and conceptual subtraction: the externalism of content. A general externalist claim is that the intentional nature of the cognitive phenomena to be explained requires that the specifications of the contents involved must make reference to the environment of the subject. This is not only because intentional properties do not in general supervene on the states of the organism alone[12], but also because intentional phenomena can only be specified, explained, and understood in terms of their directedness toward the external world and the potential to interact with it. For example, it seems very likely that a means of specification must include some way of representing the spatial environment of the subject if it is to be able to express and explain spatial NCCs and their inter-relations, as used in the construction of cognitive maps.

One might question this conclusion by noting that conceptual content is intentional, yet we specify such contents via non-embedded means (linguistic use). One reason why we can get by with non-embedded specifications for conceptual content, but not for non-conceptual content, will be given in section 5.3.2, below.

But there is also reason to believe that we can't in general get by with non-embedded specifications, even in the purely conceptual case. If externalist positions such as those expressed in [Burge, 1982] and [Putnam, 1975] are correct, then there is no way that a sentence of English language on Earth could specify even the conceptual contents entertained by our Twins on Twin-Earth. So, a fortiori, linguistic use could not specify Twin-Earth NCCs.[13] In order for the CS method to avoid this limitation, it must be the

[12] Martin Davies [Davies, 1991] gives some examples of non-conceptual, perceptual contents that do not supervene on the internal state of the organism experiencing that content; the important differences are in the way the organism is embedded in the environment, and thus the environment can be expected to play a major role, beyond the one of specifying truth-conditions, in the specification of non-conceptual contents.

[13] Note that arguments merely to the effect that symbols must be grounded, that there must be some environment in order for an agent's states to have any content at all [Harnad, 1990], do not in themselves argue against non-embedded specifications. It is only when one claims that external conditions partially individuate a content that one can put forward this kind of externalist argument for embedded specifications of content. The argument that symbols must be grounded is compatible with an internalist individuation of contents, even though it demands that such contents can only exist in the context of

case that one can subtract from an Earthly conceptual content to yield a Twin-Earthly NCC. This seems possible only if the NCCs of Earth and Twin-Earth are the same, i.e., if externalist arguments apply only at the conceptual level. Yet there are those (e.g., Davies [Davies, 1991]), who would maintain that even (some) non-conceptual contents are external. If so, we have yet another reason to reject non-embedded means of specification of NCCs. Perhaps, then, it is time we turned to embedded alternatives.

5.3.2 Embedded indexicals

One such alternative means of specification is suggested by the discussion at the end of section 5.2. There, descriptive linguistic use was rejected as a means of NCC specification because it cannot accommodate contents that have constitutive connections to perception and action. And indexical linguistic use was rejected because it alone could not specify content, but rather it must be supplemented by an environment within which language can function.

But we actually do specify contents via indexical linguistic use; we might explain RC's behaviour by saying "He started cleaning up the mess because he realised that he himself was the one making a mess", which includes a content specification by means of indexical linguistic use (the reflexive "he himself"). So either we don't need to appeal to an environment when specifying contents via indexical linguistic use, or appeal to such an environment is possible, and effortless.

It is true that we do need to appeal to the environment with such specifications. One has to know which subject is in question in order to fully grasp the significance of their first-person thoughts. This is clearer in the case of demonstratives: "He thought $< \mathit{that} >$ is a doorway" will explain why a subject ran into a false stage door only if one understands, inter alia, that the $< \mathit{that} >$ refers to the false door. Just as attributions such as "The content of the agent's visual perception was this" are effective, if at all, only when the speaking theorist and hearing theorist share the same environment (a condition that cannot, in general, be expected to fulfilled in the practising of cognitive science[14]), the

an environment toward which they are directed.

[14] But see section 5.3.5, "Self-instantiation".

specification of non-conceptual contents (indexically or otherwise) will have to recreate this context by invoking some detailed description of the agent's environment.

But if so, doesn't this just show that we already take embeddedness seriously, and make implicit appeals to a subject's environment, when we employ conventional indexical linguistic use specifications? Yes; however, the simplicity of the task of including the world in such cases is a consequence of the systematicity of the conceptual, linguistically specifiable contents involved. Such contents have conceptually elegant rules of world-involvement: e.g., "a use of the first-person mode of thought refers to the person who is using it". Once that conceptual simplicity is absent, as in the case of non-systematic NCCs, the rules for world-involvement become fragmented, non-systematic, and ad hoc, and thus demand more effort for their specification, as well as the specification of the environment in which they function.

Can the specification of such non-systematic indexicals proceed by means of linguistic use? Can there really be a term that has associated with it the world-involving function appropriate for an NCC? Technically speaking, I suppose so, if we can specify such contents at all. For once one had some theoretical grasp of the function in question, one could simply introduce a term that had that function as its indexical function. But the point is that one would have to have some way of theoretically grasping that function in the first place, since the function will not be one with which we are already familiar in our everyday use of language. Even if, strictly speaking, indexical linguistic use is possible, its possibility is contingent upon that for which I am arguing: an alternative to linguistic use specifications.

The challenge of NCC specification via indexical linguistic use will not primarily be a matter of choosing the right (non-systematic, non-linguistic) indexicals, but mainly a matter of specifying, in the appropriate ways, the subject, its context, and the relations between them. A large part of the work in developing a means of specification for NCCs will be formalising the practice of highlighting certain aspects of the subject/environment system so that a particular, non-conceptual way of representing that situation is indicated. And even the task of choosing the right indexicals will require some sophisticated way of relating the subject to its environment.

But embeddedness requires embodiment; because embedded indexical specifications must be embedded, they must take embodiment seriously. This is not principally because in order for one to be able to make reference to the relations between a subject and its environment, one must think of the subject as having a position in that environment. One may think of a subject as positioned, and yet still disembodied. Rather, the principal reason is this: in order to understand how the highlighted environmental factors play a role in fixing the content, one must have some understanding of (at least) the perceptual and motor capabilities of the system. It is for these reasons, and because one must specify the non-systematic indexical functions involved in grasping various NCCs, that embedded indexical specification must make reference to the underlying, non-intentional characterisation of a system.

5.3.3 Content realisation

The last two means of specification to be considered here, content realisation (CR) and ability instantiation (AbI), are both, unlike those before, non-conceptual specifications in that they do not express, but rather refer to, the content to be specified, and therefore employ concepts without requiring the organism to possess those concepts in order to entertain the content so specified. In the case of CR, this reference is achieved by mentioning a set of perceptual, computational, and/or robotic states and/or abilities that realise the possession of that content in a particular case or set of cases.

As mentioned before, specifications must indicate at least one content, at most one content, and must indicate contents canonically.

At Least One Content: Realisation

By requiring that the referenced states indicate at least one content, the first condition entails, in the context of Content Realisation specifications, that the states mentioned in the CR specification must realise a content; they must be sufficient for the possession of a content. One could imagine a weaker form of state- or ability-based specification, in which the states would not have to realise the content they specify, but would instead merely suggest to the theorist the content to be specified, with no accompanying metaphysical claim that those states specify the content because they realise it. But if the metaphysical relationship is abandoned, what relationship is to be put in its place? How is one to know if the states on offer will succeed in suggesting the content in mind? In the absence of answers to these questions, any alternative to the metaphysical approach is precisely the kind of specification that I am trying to avoid: one that succeeds, when and if it does, without appeal to any principle (or at least not any articulated principle). A scientific psychology requires more rigour than such a means could currently provide; it seems that such rigour could only be provided, if ever, by a means of specification informed by a theory of "suggestion" itself, i.e., a near-complete scientific psychology. Scientific psychology would have to be completed before it could proceed.

Note that there is nothing in the CR approach that precludes an externalist individuation of content. It might be true that individualistic properties alone do not determine some or all contents (although not necessarily for the reasons given in, e.g., [Putnam, 1975] and [Burge, 1982]), but this just means that the states used to specify such contents will themselves have to be externalistically individuated. This is not in itself a difficulty (*pace* [Fodor, 1981]), since there are several examples in cognitive science of such an embedded notion of state or ability.

It is important to the proper understanding of CR specification that one note that although sufficiency of the states for the specified content is required, necessity of the former for the latter is not. That is, the specification does not have to provide or even invite a reduction of the content it specifies. To specify a content by mentioning one realisation of that content is not to indicate the physical type that constitutes possession of that content in general. Indeed, CR specifications do not even require that a reduction of the content to a non-intentional vocabulary is possible (which is just as well, since there are good reasons to believe that such reductions are not possible). Conversely, the fact that there might be infinitely many other physical configurations, that fall under no nomologically-governable physical type, that are also realisations of the given content, counts not one whit against the ability of the particular realisation mentioned to pick out, clearly and distinctly, the content in question. Consider how one might indicate to someone a particular economic phenomenon by describing a particular manifestation of that phenomenon, in terms of a particular currency, set of countries, etc.

At Most One Content: Holism

In order for the states referenced in a CR specification to specify a content, they must not only realise the content in question; the second condition mentioned above, together with a simplistic notion of realisation-based specification entails that the specifying abilities must realise only that content. This creates a problem for CR specification, even if we take on board the idea offered in section 5.3.1: one does not have to specify a unique content, only a set of contents that share the property P to which one is appealing in the explanation. The problem is that CR specification appears to be at odds with the fact that content is holistic: contents come in groups, so any abilities that are sufficient for one content are going to be sufficient for others as well. Thus, it seems that CR specifications might have difficulty respecting the second condition.

One should not reply to this objection by denying the holism of content, even non-conceptual content. For even if NCC does not meet the Generality Constraint completely, some more limited form of holism may be required if we are to speak of content at all. Rather, one should reply by pointing out that to make the above objection to CR specification is to misunderstand the holistic nature of content. It is a holism that has to do with capacities. That is, while it may be true that one cannot acquire the capacities for entertaining a content (C) without thereby acquiring the capacities to entertain many others (S), it is not thereby true that that if one entertains C one must entertain all of those related contents in S. Thus, a realization may specify only one content, even if the capacity to entertain that content entails the capacity, or even the disposition, to entertain others.

Canonicality

Even if CR specifications indicate one and only one content, it must be ensured that they do so canonically. That is, they must avoid, e.g., being like the linguistic use specification mentioned before: "the content toward which the subject took the belief attitude exactly 10.3 seconds ago". One might think that CR specifications cannot specify contents canonically, since canonical specifications must invoke the essential properties of the content, while CR specifications proceed by mentioning a particular realisation of that content, which might be thought to be only contingently related to the content.

But a requirement for such strict necessity seems to be too stringent. Consider linguistic use specifications. Although some (e.g., [Morris, 1992, ch 13]) might think there is a necessary connection between a content and a word that expresses it, even they must concede that the relation between the content and the sounds or marks that instantiate the word that has that content is contingent. So if such marks are sufficient for canonical specification, then it seems possible that other entities contingently related to the content, such as one of its realisations, could also be sufficient for canonical specification.

To make good this analogy between marks and particular realising states, there needs to be more to CR specifications than mere realising states, just as there is more to linguistic use specifications than mere marks or sounds. In the case of linguistic use, there is a practical capacity, on the part of the theorist, to relate these arbitrary marks to the contents they express. This practical capacity is part of being a member of a linguistic community, and is acquired through exposure to the norms that the community applies to the sounds and marks that the language comprises. So it would seem that CR specifications, if they are to be canonical, must rely on some practical capacity for relating particular realising states to their general forms, the forms which are essential to any state that realises the content in question. And this capacity might have to be acquired through familiarity and practical interaction with the system in question.[15] Note that one would not have to develop such a capacity for every system to be explained, but only for the system or systems that one wishes to use for the purposes of CR specification of content.

However, the more that one models CR specification on the case of linguistic use specification, the more one runs the risk of limiting CR to conceptual contents. It could very well be that the requirement of a public, practical capacity to understand others is what restricts linguistic use to conceptual contents. If so, one might worry that the requirement for a practical capacity to understand the canonical realising system in CR specifications might likewise limit such specifications to the conceptual case. Although this worry cannot be dispelled entirely here, it should be pointed out that there is a stronger link between a content and one of its realising states ("intrinsic intentionality") than the relation between a content and the arbitrary properties of one of the symbols that convention and practice have associated with that content ("derived intentionality").

[15] See section 5.3.5, "Self-instantiation", for further discussion of how this might be possible.

In fact, once the general parameters of the specifying system have been determined, there might be a necessary relationship between a state, given that it is bounded by those parameters, and the content it realises. Therefore, canonical specification may be possible without relying on practical capacities that might restrict one to conceptual contents. But perhaps this is just optimism; at present the issue is unresolved.

The limitations of both non-embodied (linguistic use, CS) means of specification, as well as of those that merely mention embodiment (EI, CR), might have a common cause. Specifically, it might be that the only way canonically to specify NCCs is via an explicit demonstration or actual instantiation of the idealised robotic and computational abilities involved in entertaining that content. Attempts to specify an NCC in a linguistic use manner fail to indicate (to any theorist seeking to understand the agent) the correct content and therefore leave certain connections to perception, action, and other contents inexplicable; perhaps merely mentioning the abilities must also fail, for similar reasons. Practical, canonical specification of an NCC might require the actual instantiated presence of an ability, rather than the conceptual idea of that ability. There are two possible ways that the abilities could be instantiated: external to the theorist, in some apparatus (external ability instantiation, or EAI); or within the theorist/environment system itself (self-instantiation, or SI).

5.3.4 External ability instantiation

There are practical reasons why actual instantiation of specifying abilities, as opposed to mere reference to them, might be required. The demands for embeddedness in content specification seem to call for a means of specification that not only allows one to represent explicitly the spatio-temporal relations between the system being modelled and its environment; it must also itself be a concrete system that persists through time and possesses computational abilities. As argued before (in the discussion of the possibility of embedded indexical (EI) specifications in section 5.3.2), the context of the subject will have to be reconstructed if we are to be able to specify (at least some) NCCs; this in turn requires a specification of at least the perceptual and motor capabilities of the system. This specification should be achieved, at least in part, via a judicious choice of the non-semantic properties of the specification formalism itself. It should have an active, computational format, rather than the static format of axioms and theorems on a printed page. The complexity of a fine-grained static formalism with spatio-temporal parameters (e.g. "$Believes[robot_3, time_1,$ *that is an obstacle*, the chair at location (x, y, z)]") would be prohibitive; instead, a computer simulation (of the interactions of the various axioms of the theory and boundary conditions of the situation) would make explanation and prediction more tractable than if one were to use non-instantiated, referential, 'manual' analysis. In order to understand which content is involved in a particular situation, a theorist could look at how an instantiated system responds to different counter-factual contingencies, could monitor the evolution of the current state forward or backward in time, etc. This would not be feasible with a non-instantiated means of specification.

But there is a more theoretical reason why an external, active instantiation of some abilities that realise a non-conceptual content might be necessary for the canonical specification of that content. EAI specification can be seen as a response to the worries, just expressed at the end of 5.3.3, concerning the canonicality of content realisation (CR) specifications. Perhaps one can specify contents, as CR aims to do, in terms of the states and abilities which realise them, but the pre-requisite practical ability, on the part of the theorist, to move from realizing abilities to specified contents might require an active presence of those abilities, with which a theorist can interact, not just reference to them. The observable temporally-extended action of the computational formalism (and its interaction with its environment) might be the only way canonically to specify certain NCCs, given their resistance to specification by standard means. Some phenomena may only be explicable through the use of models; perhaps only via models in which, e.g., actual time and space are used to represent the temporal and spatial aspects of the modelled system, as opposed to formalisms that represent those aspects with something else: a written variable or spacing on a page. It seems likely that in order to be able to specify NCCs and their inter-relationships, one will have to choose representations for them in such a way that there is a non-arbitrary relationship between the non-semantic properties of the representations and the contents to which they refer: the non-semantic properties will assist directly in specifying the content.

This approach (and others presented here, inasmuch as they are concerned with the question "what is a canonical specification?") places an emphasis on the theorist's own embodiment, with the notion of a theorist's psychology that such embodiment implies. Canonical specification cannot proceed independently of the cognitive make-up and limitations of the theorist using that specification; rather, what counts as a sufficient specification or, more generally, explanation, will depend on the conditions under which the theorist's abilities to grasp contents may be exercised. One tentative proposal is that our psychologies as theorists are such that we will only have canonical NCC specification when we employ actual instantiations of that content.

5.3.5 Self-instantiation

There are two different proposals to be considered in this section, although both are similar in several ways, including their speculative nature and science-fiction feel. Let that serve as a qualification.

The first form of self-instantiation continues the realisation thread under consideration in sections 5.3.3 and 5.3.4. There a worry was expressed: that the practical capacity required to move from realisations of contents to the contents themselves will have to be similar to the practical capacity for language to such an extent that only linguistic, conceptual contents can be so specified. Despite the observation that the relationship between realisations and contents is less arbitrary than that between sounds or marks and linguistic contents, the discussion in 5.3.4 suggests that the practical capacity may have to be an almost social kind of interaction with the specifying system if it is to provide canonical specifications of NCCs. This fuels the worry.

Perhaps the instantiation that provides canonical specification should be something more intimately known, thus avoiding the need for interactive, social capacities. Perhaps the instantiation should be the theorist itself. This suggestion is supported by the view that a state characterized by content that presents the world as less than objective is itself less than objective, in a particular sense. That sense is: a state is objective to the extent that one subject can understand another subject as being in that state without having to have properties in common with that other (compare [Nagel, 1986]). A state which one can only understand by actually being similar to someone in that state is a subjective state. Section 3.6 of chapter 3 argues that conceptual content is objectual, thus the extent to which a content is non-conceptual is the extent to which it is non-objectual, and hence non-objective. If the particular notion of objectivity introduced here is a part of the notion of objectivity already employed in this thesis, one may conclude that the less[16] conceptual a content, the greater the extent to which a theorist (wishing to specify that state canonically) must share properties with another who is entertaining that content. Thus, we have an argument from the non-conceptuality of a content to the need for a self-instantiation specification of that content. Another way of seeing the point is this: to specify a content, one must always self-instantiate. But because conceptual content is individuated with respect to judgement alone (chapter 3, section 2.1), one need only be able to alter oneself doxastically in order to be able to instantiate, and thus specify, conceptual content. Conversely, since non-conceptual content is individuated with respect to more than just role in judgement, specifying non-conceptual content in general requires one to be able to alter oneself in ways other than just the concepts one grasps, beliefs one holds, etc. Thus, non-conceptual content requires some more robust form of self-instantiation.

The situation I have in mind is not the use of some private "inner pointing", against which Wittgenstein argued. Rather, imagine a (possibly not-too-futuristic, given recent advances in, e.g., neurophysiological imaging techniques) situation in which the theorist learns the relation between publicly observable states/abilities and contents for his or her own particular case. The theorist's non-intentional state at any given time will be directly observable, and the theorist will have a privileged (though not necessarily infallible) acquaintance with the corresponding intentional state. This combination of extro- and intro-spection may permit the development of a practical capacity for the theorist to mention his or her own physical states to specify contents that would otherwise be ineffable. If one further assumes that the realisations of theorists' contents do not differ dramatically from each other, there will be the possibility of theoretical, scientific communication concerning NCCs.

This means of specification is not, strictly speaking, one of state/ability instantiation,

[16] That conceptuality admits of degree is explained in section 3.7 of chapter 3.

since the specifications themselves could very well be references to or mentions of the states of the theorist that realise the content in question. But since the development of this capacity for such states to specify contents canonically requires a period in which the theorist actually instantiates (perhaps some "basis" subset of) the contents to be specified, it seems appropriate to mention this method under the "instantiation" rubric. Any of the instantiation methods could have an initial period of specification via instantiation, during which technical terms are introduced to refer to the contents so specified. But use of such terms for specification would still be a case of instantiation specification, since the norms of use of such terms is governed by their means of introduction.

However, the second form of self-instantiation specification is more directly a case of instantiation. It also has more of the feel of the "inner pointing" which Wittgenstein argued against, yet with a grounded twist that might allow it to avoid coming under the purview of his private language arguments. The idea here is to cut out the middle man, by altering the environment of the theorist (to which one wishes to communicate a content) such that the theorist actually takes an attitude toward that content.

This can be seen as playing the same role for Embedded Indexical specification, in terms of grappling with the constraints placed on canonicality by the nature of the theorist's cognitive abilities, that External Ability Instantiation and the first form of Self-Instantiation played for Content Realisation. For example, in a typical Embedded Indexical specification, one might say "the infant sees the wall like $< \textit{this} >$", followed by a description of the infant's environment, its position and orientation within that environment, and its sensori-motor abilities. The analogous move to that made before, then, is to claim that this referential approach is not sufficient for canonical specification, a more instantiated approach is required. The move would claim that any success for the EI method would be due to the theorist being able to imagine the situation from the infant's point of view. But our imaginations are notoriously limited; why not actually have an externally-prompted experience with the same content as the one to be specified?

Clearly, it would be too awkward (at present) for one to manipulate a theorist's environment to the extent necessary for such specifications. But we cannot rule out the possibility that technology (e.g., virtual reality) could be of assistance here, if or when it is developed.

Nevertheless, there are obvious potential difficulties: given the holistic nature of content, could an adult theorist, no matter how his or her environment is manipulated, really see the world the way an infant does? Even if one believes, as is surely the case, that adults entertain a wide range of non-conceptual contents, is it plausible that they are the same contents that are entertained by an infant? A bat?

Rather, it seems that if canonical specification requires that close of a link between theorist and subject, then we are severely limited in our capacities to understand each other from a scientific viewpoint. I choose to interpret this as a strike against such a strong notion of the requirements for canonical specification, rather than against the prospects for a scientific psychology.

5.4 Embodiment and computation

No matter which (if any) of the types of alternative specification actually turn out to be successful, it seems clear that NCC specification requires appeal to the spatio-temporal relations between the system being modelled and its environment (embeddedness); for this (recall the end of 5.3.2) and other reasons, then, such specification also requires either reference to, or the instantiation of, (some of) a content-exercising system's intrinsic non-intentional properties (embodiment).

It is natural to look to computation and robotics to provide ways of characterising and thinking about the functionally relevant aspects of the system's embodiment and its environment. But there are two thoughts that might give one pause.

First, it is arguable that computational phenomena are themselves intentional. Computational states are typically representational, they are about things, they carry their own form of (sub-personal) content. So one might wonder how computational notions could provide the characterisations of non-intentional states required for NCC specifications. For indeed the embodiment and environment of the system must be characterised as non-intentional (or at least non-contentful) if an infinite regress of content specifications is to be avoided. But computational analyses specialise at coming up with elucidating, un-interpreted (if not downright non-intentional) ways of characterising intentional systems. Of course, computational states are intentional, are about something; but viewing them as, say, Turing Machine quadruples is to highlight their merely causal properties, and to ignore their semantics. Perhaps the value of this kind of analysis has been overemphasised, or misunderstood; I certainly don't think that a complete understanding of computation will be primarily formal and non-intentional (see chapter 6). Nevertheless, such characterisations do have their place, and they might be ideal candidates for capturing the embeddedness and embodiment of systems for the purpose of content specification.

On the other hand, some might think of computation as a world-independent, abstract notion, not the kind of thing that could square well with the requirements of embodiment and embeddedness at all (again, see chapter 6). All that can be said here is that there are reasons, discussed elsewhere, for rejecting this disembodied, asemantical view of computation (see, e.g., [Smith, 1991]). In fact, one could put the force of the issue the other way: given that content specifications must be embodied and embedded, if cognitive science is going to understand representational content in terms of computation, we had better develop our computational notions accordingly, rejecting the formal for the embodied and embedded.

One way that non-conceptual content and computation relate, then, can be captured with the motto: "do not ask what your formalism can do for your robot; ask what your robot can do for your formalism". That is to say, it seems that in order to specify contents and their inter-relations, a means of content specification for an NCC-involving cognitive science will require concepts and insights from a theory of computation (especially robotic and perceptual computation). Further, such a formalism might require not only concepts and insights, but instances of computational phenomena.[17]

5.5 Conclusion

The existence of non-conceptual content (NCC) places several demands on any cognitive science theory that wishes to address the full range of human cognitive behaviour. I

[17] Thus, my discussion here does not concentrate on computation via the claim that cognition is computation (although that equation might be a consequence of the concerns here); rather, the emphasis is that computation is a (and perhaps the only) formal means of specifying otherwise ineffable contents.

have argued that the way to answer these demands is to take embodiment seriously, by establishing a close connection between NCC and computational/robotic abilities. I argued that we need an alternate means of content specification that can, unlike the standard method of linguistic use ("that" clauses), canonically specify NCCs. I suggested that a worked-out means of specifying computational and robotic abilities might go a long way to meeting these requirements. The demands that must be met before a fully worked-out means of specifying NCCs can be given are considerable, but they should not discourage: an emphasis on NCC not only constrains, but also liberates, in that it allows psychologists to direct their energies toward explaining cognitive phenomena which have to be ignored from within a conceptualist approach, since the phenomena essentially involve contents which are non-conceptual: cognitive phylogeny, conceptual development, perception, learning, and action.

CHAPTER 6

The ontological status of computational states

6.1 Introduction

As said before: an account of intentionality needs to be naturalized; cognitive science aims to do this by appealing to content-based explanations, and to representational vehicles which can be seen to carry those contents. In particular, cognitive science appeals to the notion of *computational* representational vehicles (see 7, section 7.1). Furthermore, the primary claim of chapter 5 is that specification of non-conceptual content requires use of computational notions.

A specific worry about our current understanding of computation arises out of the observation that our formal notions of computation, such as those expressed in the formalisms of Turing Machines and recursive function theory, seem so abstract as to deem computational any physically realizable system. The worry focusses on the lack of utility of a concept of computation that is as universally applicable as physical realization. If any physical system can be characterized as computational, how can it be interesting that a particular system is computational? How can the fact that that system is computational be explanatory? In particular, how can the notion of computation be used to explain cognition, to distinguish thinking beings from mere inert matter? It seems we need a more restricted notion of computation.

Both Putnam [Putnam, 1988, pp 95-96; 121-125] and Searle ([Searle, 1990], [Searle, 1992, ch 9]) have presented arguments for the claim that computational states are universally realizable, in the sense that we could interpret any physical system as instantiating any computational characterization. They both argue that this has dire consequences for the computational view of the brain and mind that is a working hypothesis in cognitive science. For example, Searle puts it this way: just as one can argue (via the Chinese Room argument) that semantics is not intrinsic to syntax, so also can one argue that syntax itself is not even intrinsic to physics [Searle, 1992, p 210]. But whereas Searle admits [Searle, 1992, p 209] that the threat of universal realizability could be avoided if our notion of computation is modified to include causal and counterfactual notions (implying that these are lacking at present), Putnam thinks that the universality, and hence vacuity, of the notion of computation remains, even if one requires computational state transitions to be causal.

In the following, I analyze Putnam's argument and find it inadequate, because, inter alia, it employs a notion of causation that is too weak. Therefore, the fact that a particular system realizes a particular automaton is not a vacuous one, and is often explanatory. But also I claim that computation would not necessarily be an explanatorily vacuous notion even if it were universally realizable. Even if it is possible to characterize all systems computationally, in many cases it will not be explanatory to do so, but sometimes it will. Thus, claims such as "the brain is a computer executing program P" are not meaningless or incoherent, as Putnam would have us believe. Instead, such claims are quite contentious, and can be true – or false.

But before turning to Putnam, further consideration of Searle's position is required. Despite what I said two paragraphs before, I do not mean to suggest that Searle thinks all is rosy about the ontological status of computational states. He says [Searle, 1992, p 209] that perhaps one can't interpret any physical system to be any computer, but that doesn't matter, since the *real* problem with computation is that it involves a notion of interpretation in the first place. This makes computation observer-relative, and therefore unsuitable as a foundation for cognitive science. I think there are two ways in which Searle thinks that even a causal notion of computation is observer-relative, but I think neither should worry anyone who wishes to found an understanding of the mind on computation.

6.1.1 Multiple interpretations

First, there is an objection to (even a causal notion of) computation that arose in personal discussions that I have had with Searle (but of course, he is not *committed* to the views I ascribe to him here). I believe Searle would consent to the following: if one adopts a causal notion of computation, then every system will *not* realize every computation, but every system *will* realize multiple (perhaps infinitely many) computations simultaneously.

I agree with all that; so far so good. The disagreement between Searle and me comes next: he thinks that this realization of a multitude of computational descriptions is still a problem for a computational foundation for cognitive science. Why? Presumably because he thinks cognitive science requires that there be a unique computational description for a system that is to be explained. But to single out a particular computational characterization in such a way is to make cognitive science observer-relative: one could have been just as justified in choosing a *different* computational characterization for the *same* system.

But cognitive science doesn't require that there be a unique computational description for a system. Consider a cognitive science that uses computation in the following reductive sense: mental states *are* computational states. On this view, there are a host of laws of the form: anything in computational state C (individuated by appealing to a computational description) is thereby in mental state M. (I suspect that *identity* is too strong to be the right relation between computational and mental states, but if Searle's objection fails for even this extreme form of computationalism, it will a fortiori for weaker positions.) Presumably, Searle's thought is this: since there are multiple computational characterizations of a system, it will follow that the antecedents of more than one of these laws will be satisfied, and therefore there will be some indeterminacy as to which of the several mental states mentioned in the consequents of the activated laws is the *real* mental state of the system. This indeterminacy can only be resolved by arbitrarily choosing to employ one computaional description over the others. Thus, mental states would be unacceptably observer-relative.

Some might not think that this result would be objectionable; but I share Searle's desire to avoid such observer-relativism of the mental. Fortunately, such indeterminacy doesn't follow from the fact that any system realizes a multitude of computational descriptions.

It does not follow for at least two reasons:

> Clearly, not *every* computational state will appear in the left hand side of one of these laws; every physical system can be correctly characterized as the one state finite automaton, so nothing should have any mental states in virtue of realizing *that* computational description. In fact, it might be that out of all the computational descriptions that a given system realizes, *only one* will appear in the antecedent of a computational/psychological bridging law; or, it might be that all the computational descriptions appear on the left hand side of the *same* law. In such cases, there would be no multiple assignment of mental states, no indeterminacy, and thus no observer-relativity.

> Even if more than one of the computational descriptions appears on the left hand side of a bridge law, and even if they appear in *different* laws, the multiple mental states so assigned might not be incompatible, either because the multiple mental states are hierarchically related (e.g. I'm happy, and I'm happy that today is Friday; no indeterminism there) or because the mental states just simply *can* be possessed at the same time (e.g. I'm happy that today is Friday, and I believe that today is Friday).

6.1.2 Interpretability

The second reason why one might think that computation is observer-relative, the one Searle gives in his book, is this:

> We can't, on the one hand, say that anything is a digital computer if we can assign a syntax to it, and then suppose that there is a factual question intrinsic to its physical operation whether or not a natural system such as the brain is a digital computer. [Searle, 1992, p 209-210]

This brings us to issues of realism and instrumentalism in science that are too large to be addressed in this digression, but I have a quick reply. That an object is *interpreted* by someone as being C is a deeply observer-relative fact; that an object is *interpretable* by someone as being C need not be observer-relative, if enough constraints are put on the conditons of interpretation. Whether or not a particular phenomenon is interpretable by us in a certain way does not just depend on *us*; it also depends on the phenomenon. Many, many things can be interpreted by us as being, say, a particular 25-state Turing Machine. But *vastly* many more will not be so interpretable. That suggests that there is something that those interpretable things have in common, something objective, even though that objective commonality happens to have a convenient expression in terms of our abilities to interpret.[1]

6.2 Putnam's argument for the universal realizability of finite automata

Putnam has provided a meticulous and concrete expression of the claim that computation is so abstract as to be vacuous. His "theorem", if its complex derivation is sound, establishes that "every ordinary open system is a realization of every abstract finite automaton." In order to establish his conclusion, Putnam appeals to two physical principles:

> *The Principle of Continuity.* The electromagnetic and gravitational fields are continuous, except possibly at a finite or denumerably infinite set of points. (Since we assume that the only sources of fields are particles and that there are singularities only at point particles, this has the status of a physical law.)
>
> *The Principle of Noncyclical Behavior.* The system S is in different maximal states at different times. This principle will hold true of all systems that can "see" (are not shielded from electromagnetic and gravitational signals from) a clock. Since there are natural clocks from which no ordinary open system is shielded, all such systems satisfy this principle. (N.B.: It is not assumed that *this* principle has the status of a physical law; it is simply assumed that it is in fact true of all ordinary macroscopic open systems.) [Putnam, 1988, p. 121]

The Principle of Continuity claims that the electrical and gravitational fields are continuous; the Principle of Noncyclical Behavior states that every system is in different states at different times. The first principle I will not dispute, other than to point out that as Putnam admits [Putnam, 1988, p. 121], the Principle of Continuity assumes classical, as opposed to quantum, physics. The impact of this assumption on the success of Putnam's argument I leave to those who can speak on such matters with authority.

The second principle, however, is more problematic, as is the way that Putnam attempts to employ it. Briefly, the only way Putnam can guarantee the truth of the second principle

[1] Further, on the broad notion of "observer-relative" that Searle's discussion requires, don't our *other* scientific physical properties (e.g., biological ones) involve, at root, some notion of interpretation? If *they* are observer-relative, then what's so wrong with being observer-relative?

is for him to individuate states by their absolute position in time; but this prevents him from using the principle in the way he intends: to demarcate states that are causally related in such a way as to realize a particular finite automaton (cf. section 6.5, below).

Putnam's argument proceeds as follows. He sees it as sufficient to show how any physical system can realize some arbitrary finite automaton, such as one that goes through "the following sequence of states in the interval (in terms of 'machine time') that we wish to simulate in real time: $ABABABA$" [Putnam, 1988, p. 122]. The goal is to come up with a definition, in terms of the physical properties of an arbitrary system S, of the states A and B such that the system goes through the sequence of states $ABABABA$ in a particular time interval. Let $t_1, t_2, ..., t_7$ be the times corresponding to the beginning of each of these automata states, with t_8 being the time of the end of the last state. Let s_i be the region of physical state space that S occupies between t_i and $t_i + 1$. The definitions for A and B in this particular case (and therefore, in principle, in general) are easy to state: $A = s_1 \vee s_3 \vee s_5 \vee s_7$ (i.e., the system is in computational state A if its physical state lies in any of the the parts of state space denoted by s_1, s_3, s_5, and s_7); $B = s_2 \vee s_4 \vee s_6$. This will entail that S is in states A and B at the right times to result in the sequence $ABABABA$ for the temporal interval in question.

We can see immediately an example of Putnam's need to appeal to his physical principles. Without the Principle of Noncyclical Behavior, one cannot assume that the s_i will be disjoint, and if that is so, then some of the conditions sufficient for A might turn out to be sufficient for B. For example, if s_2 were not disjoint from s_3, then there would be at least one point in state space that is in both s_2 and s_3, implying that when the system was in that physical state, it would also be in both computational states A and B. This would yield an ambiguous interpretation function from the physical states of S to the computational states of S, whereas automata states are exclusive.[2]

[2] Presumably, even those wishing to establish the universal realizability of computation would agree that ambiguous (one-to-many) interpretation functions could not provide an adequate notion of computation. Otherwise, their claim is trivially established: any system realizes any finite automaton because every physical state can be mapped to every computational state, even under the same interpretation. At any rate, those wishing to defend computation as non-vacuous merely have to stipulate that computational properties supervene (at least) on physical ones (i.e., if you change the computational state, you must

So the stakes for the s_i being disjoint are high. If they are not, Putnam can't ensure that he will always be able to construct a proper, non-ambiguous interpretation function from physical to computational states. That's where the Principle of Noncyclical Behavior comes in: the disjointness of the s_i follows directly from the purportedly noncyclical behaviour of S. If S never makes transitions to states in which it has been previously, then there is no way that the temporally disjoint s_i (which are just time-slices of S) could fail to be disjoint in state space. Thus the stakes are moved from the disjointness claim to the second principle which supports it. But, as I will argue below (in section 6.5), Putnam gives us no reason to believe that systems can never be in the same state twice.

6.3 Is computation essentially causal?

Ignoring, for now, the problems with the disjointness of the s_i, the only thing then left for Putnam to show is that the sequence of state transitions is causal; that the fact that the system is in state A (and receives the input that it does at that time; this is discussed in section 6.6 below) *causes* the system to go into state B (and emit the outputs that it does). Putnam has to show that his arbitrary computational interpretations of a state are causal; otherwise (as Searle admits) one could prevent universality by only considering the causal characterizations to be the ones that are truly computational.

Some might deny that causal connectedness is an essential property of computational states. Turing Machines themselves, after all, are completely formal; they are abstractions, and are therefore not the kinds of things that can have internal causal structure. However: even if the formal abstractions themselves are not causal, it is a mistake to think that there can be no causal requirements which a physical system must meet in order to be a realization of a formal abstraction. The very fact that they are called Turing *Machines* suggests that the transitions between the realizing states must be mechanizable, or at least causal.

Furthermore, consider an animated display of a Turing Machine on a computer screen. Since, ex hypothesi, there is a one-to-one correspondence between the states of the display screen and the states of some Turing Machine, Searle and Putnam would apparently claim

change the physical state somehow) in order to reject this extreme form of universality.

that the screen realizes the Turing Machine, if anything does. But it seems clear that we would say that the screen *depicts* a Turing Machine, but is not itself one. One reason why we would deny it computational status is because the state of the screen that corresponds, in the putative interpretation function, to a computational state A does not produce, as a causal effect, the screen state that corresponds to the successor computational state B, even though the Turing Machine depicted does make a transition from state A to state B. Computational states must be able to *cause* other computational states to come about.[3]

But those arguments only establish that we do, in fact, take causation to be essential to computation. But why *should* we, other than to avoid the universal realizability results? One reason seems to be this: computational characterizations are not purely descriptive; they are also explanatory and predictive. In virtue of characterizing something computationally, we not only describe its past, but predict its future and explain both. The fact that our notion of computation puts some constraints on the intrinsic, causal properties of the physical systems which realize that computation allows us to use a computational characterization in order to *predict* the behaviour of that system. If there were no connection between our computational notions and causation, then we would have no reason to expect a physical system to continue to be interpretable (with a fixed interpretation function) as realizing a particular computation. Of course, one could, in an ad hoc manner, continually modify the interpretation function from physical states to computational states, so as to guarantee that the system will continue to realize a particular computation. This is, in fact, what Putnam suggests we do. But this method, unlike a truly causal understanding of computation, would not allow us to *predict* which intrinsic physical states a system will go through in the future. We can logically guarantee that any physical system will enter the computational state A in the future only by giving up all claims as to the intrinsic nature of the realization of A, and thus giving up all predictions of the behaviour of the

[3]One might agree that the screen states are not causally related, but argue that neither are the bits in screen memory, bits in RAM, or voltages. That is, yes, the screen states are mere depictions of Turing Machine states, but it is depictions *all the way down*. I disagree. There is some complicated set of CPU, memory, wires, voltages, etc. which causally realize the various Turing Machine states. Otherwise, given, *in advance*, a particular scheme of interpreting physical states as computational states, it would be a miracle, a fluke, that we could reliably get this stuff to simulate a particular Turing Machine.

system based on it being in A.

6.4 The causal efficacy of computational states

As said before, Putnam accepts that he must establish a causal connection between his constructed computational states. He argues that S being in A and having the boundary conditions that it does when it is in A *causes* S to go into state B. His argument uses the following lemma:

> *Lemma.* If we form a system S' with the same spatial boundaries as S by stipulating that the conditions *inside* the boundary are to be the conditions that obtained inside S at time t while the conditions on the boundary are to be the ones that obtained on the boundary of S at time t', where t is not equal to t' [note that this will be possible only if the spatial boundary assigned to the system S is the same at t and t'], then the resulting system will violate the Principle of Continuity. [Putnam, 1988, p 121]

The argument for causal connectedness then proceeds by claiming that given the state of the boundary of S at time t, then, by the lemma and the Principle of Continuity, the inside of S *must* change from the state it was in just before t to a state distinct from any other state it occupies in the time interval under consideration. Thus, the transitions between states are causal.

I think that Putnam's argument for the causal connectedness of his constructed computational states is unconvincing for several reasons:

1. It relies on the Principle of Continuity;

2. It relies upon the lemma, which, as I will argue in section 6.5, lacks justification, for the same reasons as does the Principle of Noncyclical Behavior and therefore his argument for the disjointness of the s_i;

3. It manages to establish causal links between the states of arbitrary physical systems only by assuming a very weak notion of causation.

Since I've already expressed some doubts concerning Putnam's continuity assumptions (1), and the lemma (2) is discussed in section 6.5, below, we can move on to Putnam's notion of causation (3).

The question is: under what construal of causation will the "connect-the-dots"-style computational descriptions that Putnam constructs entail, in general, causal relations between computational states? Putnam tells us: it is the notion of causation "that commonly obtains in mathematical physics" [Putnam, 1988, p 96]. By this, Putnam means a notion of causation that is quite weak:

> In certain respects the notion of causal connection used in mathematical physics is less reasonable than the common sense notion... If, for example, under the given boundary conditions, a system has two possible trajectories – one in which Smith drops a stone on a glass and his face twitches at the same moment, and one in which he does not drop the stone and his face does not twitch – then "Mathematically Omniscient Jones" can predict, from just the boundary conditions and the law of the system, that if Smith (the glass breaker) twitches at time t_0, then the glass breaks at time t_1; and this relation is not distinguished, in the formalism that physicists use to represent dynamic processes, from the relation between Smith's dropping the stone at t_0 and the glass breaking at t_1 [Putnam, 1988, p. 97].

This is a weak notion of causation in that the conditions, under this notion, that have to be met in order for two events to be causally related, are weaker than the conditions for our common sense notion. For example, our common sense understanding of causation would not deem Smith's twitching and the glass breaking as causally related, while Putnam's understanding would.

In order to support this notion of causation, Putnam attempts to discredit what he considers to be the main alternative: a notion of causation based on possible worlds and counter-factual conditionals:

> ...one can sum this up as follows: when we consider what would have been the case if Smith had not twitched, we keep such things fixed as that he released the stone. This means that... we consider situations in which the boundary conditions themselves (or the initial conditions, or both) are quite other than they actually are [Putnam, 1988, p. 97].[4]

Putnam's objection is that any account of causation in terms of counter-factual conditionals is dependent on a prior notion of what range of possible worlds, for each A and B,

[4]It is odd that Putnam emphasizes that the possible worlds notion of causation considers "situations in which the boundary conditions are quite other than they actually are." For the mathematical physics notion, too, must vary at least some of the boundary conditions. Otherwise, the only systems that would have different "possible trajectories" would be non-deterministic ones, yet Putnam has stated that he is focusing on the classical (hence, presumably, deterministic) case.

are to be used for the determination of whether A caused B. And the idea of a similarity metric on possible worlds is in at least as bad shape as the notion of computation which it is supposed to explicate. Putnam also claims that the notion of "possible world" itself is in dire need of explication. But if this is so, it undermines his own favoured theory of causation as well, since that theory appeals to the "possible trajectories" of a system. The difference between Putnam's notion and the counter-factual notion of causation is not that only the latter uses a notion of possibilities; it is that only the latter uses a similarity metric to determine *which* possibilities are to be considered. Putnam's notion, supposedly, considers all possibilities equally.

This is not the proper place for a detailed enquiry into the advantages and disadvantages of a possible worlds approach to causation, but a more general point can be made: at most Putnam has only showed that one's account of computation will be as universally realizable as one's account of causation. *If one sees causation everywhere, then one will see computation everywhere.* If, however, one prefers to work with a notion of causation that is more restricted, that conforms more to our common sense notion of causation (even though a full account of such a notion may be a long time in the coming), then one will be able to make sense of the idea that some physical systems instantiate a particular computational system, and some do not. I think there are good reasons for favoring, in science, a distinction between two contiguous events that are related causally (the dropping of the stone and the glass breaking), and two contiguous events whose contiguity is merely a matter of coincidence (the twitching and the glass breaking). This is precisely what causation is meant to do; a notion which doesn't do this (such as Putnam's) isn't really a notion of causation at all.[5]

[5] However, those who wish to naturalize intentionality with computation should take heed of a difficulty that Brian Smith has suggested to me in personal discussions. If our account of computation does depend on a notion of similarity of possible worlds, and if the proper account of similarity of possible worlds is itself an intentional one, then it appears that an account of *all* intentionality in computational terms would have to be circular. Perhaps computation can only help naturalize some subset of intentional phenomena?

6.5 Complexity requirements for computational interpretation

Searle seems to be aware of the fact that the physics of a system *do* constrain the possible computational ascriptions to that system when he mentions that a system must be "sufficiently complex" in order to be understood as instantiating a particular computation [Searle, 1992, p 208-209]. Putnam also realizes this; for example, he would admit that a system cannot be assigned computational state A at t_1 and B at t_2 if its physical state at t_1 is indistinguishable, in terms of its intrinsic properties, from its physical state at t_2. It's just that Putnam believes that every *ordinary* open physical system is, in fact, arbitrarily complex (i.e., can be individuated into the number of distinct states necessary to instantiate any automaton).[6]

The last reason, then, for rejecting Putnam's argument for the causal relatedness of his constructed computational states, and for rejecting his Principle of Noncyclical Behavior, centers on his claims concerning the arbitrary complexity of physical states. Specifically, the problem is the lemma mentioned before: if a system were to have the boundary of S from one time and the interior of S from a different time, it would violate the Principle of Continuity. The problems arise in his unconvincing proof:

> Proof (of the lemma): Every ordinary open system is exposed to signals from many clocks C (say, from the solar system or from things which contain atoms undergoing radioactive decay, or from the system itself if it contains such radioactive material – in which latter case the system S itself coincides with the clock C). In fact, according to physics, there are signals from C from which it is not possible to shield S (for example, gravitational signals). These signals from C may be thought of, without loss of generality, as forming an "image" of C on the surface of S. For the same reason, there are also "images" of C *inside* the boundary of S. The "image" of C at, say, $t' = 12$ may be thought of as showing a "hand at the 12 position"; while the "image" of C at, say, $t = 11$ shows a "hand at the 11 position." Thus, for these values of t and t', the system S' would have a "12 image" on its boundary and an "11 image" at an arbitrary small distance inside its boundary; but this is to say that the fields which constitute the "images" would have a discontinuity along an entire continuous area, and hence at nondenumerably many points. [Putnam, 1988, p. 121-22]

[6] Therefore, strictly speaking, Putnam is not claiming that computation is *universally* realizable, since there may be some systems that are shielded from every clock. But that alone is not enough to give the computationalist any solace, for reasons similar to those discussed in section 6.6, below. For example, anyone who wishes to claim that mental states are computational states would have to admit that not only does a stone have mental states, but it has *all* possible mental states.

Why is this not convincing? Because Putnam assumes, without justification, that the "images" on the boundary and interior of S are characteristic of the current time of the clock that generates the images. And he assumes that they are characteristic in a strong sense: the images of the signals that bombard S are dissimilar to such an extent that a system with a boundary image of t and an interior image of any t' distinct from, but arbitrarily close to, t would violate the Principle of Continuity.

Putnam obviously does not intend to use a temporally relational individuation of physical states. If he did, then he wouldn't have had to bring in the empirically questionable Principle of Noncyclical Behavior in order to *argue* that systems are in different states at different times; he could have just *stipulated* this. He must, therefore, be using a relatively intrinsic individuation of physical states. In order for the argument for the lemma to make any sense, then, it must be that one of the following is what Putnam imagines to be the case:

> All systems have "counters" that take as input the gravitational signals, radiation, etc. they receive and increment their count accordingly. This counting ability must be arbitrarily robust: there can be no limitation on how high a system is able to count if Putnam is to be able to make his claims.
>
> All clock signals explicitly (i.e., in terms of their intrinsic properties) encode their absolute position in time. Thus, systems that are bombarded by them are never in the same state twice, since they have a new input at each instant.

It seems that Putnam must take one of these views in order to claim that the "images" of a particular clock time are characteristic of that time. If they are not characteristic, then it might be that the images corresponding to two different times would be the same, and therefore, his lemma would be shown to be false. That is, no discontinuity would occur if the images of those two times were simultaneously present in the boundary and interior. And if that were the case, then Putnam hasn't shown that the system *must*, even given the boundary conditions, make the state transitions that it does. As a consequence, Putnam could not guarantee that the relations between his constructed computational states are causal, *even on his weak notion of causation.*

So he has to appeal to something like the two ideas just mentioned. But both of these options have problems. As far as the first one goes, one has to ask what physical law prevents a system from being a flip-flop? It seems very likely that there are systems that receive a steady stream of qualitatively identical input from some clock, but merely make a transition from one of two states to the other upon receipt of these signals. How could such a system be interpreted to be realizing any automaton with more than two states, without using some ambiguous interpretation function? We saw before that such a move would be of no use, since computational realists could restrict their notion of computation so as to exclude systems with ambiguous or relational interpretation functions. Some physical systems just don't have the complexity to be interpreted as having such counters.

The second option is suggested as the one that Putnam has in mind when he speaks of "the fields which constitute the images". That is, Putnam takes those parts of the gravitational and electromagnetic fields within the boundaries of a physical system to be parts of that system. It is only by making this assumption that the discontinuity of the images could result in a violation of the Principle of Continuity, since the Principle only concerns the continuity of the gravitational and electromagnetic fields.[7]

But if this is what Putnam is assuming, then it is clear why he thinks any physical system is complex enough to realize any formal automaton. It is because he is assuming that all physical systems are continuous (via the continuity of the fields and the inclusion of the fields into the physical system). This again raises an issue from section 6.2: is it wise for Putnam to rest his philosophical points on a particular physics which ignores the discrete (quantum) nature of physical systems?

However, even if we grant continuity, and the existence of clocks which explicitly encode their time (perhaps the background radiation is an electromagnetic example; I can't imagine what Putnam has in mind for a gravitational equivalent), and the *possibility* of systems whose internal states (including the fields) reflect this temporal encoding, that does not mean that all or even any actual physical systems do, in fact, contain such images. The effects of two different clocks can cancel one another out (consider a physical system

[7] My thanks to a member of the audience at the G.H. von Wright Research Seminar reading of this paper, who pointed this out to me.

midway between two clocks that emit complimentary signals); signals can be disturbed, distorted, blocked; they can decay; qualitatively distinct signals might have identical effects on a system; etc. Surely Putnam doesn't want his argument to depend on issues as empirically contingent and contentious as *these*?

Since it seems that Putnam can't, without further justification, appeal to the lemma, he has given us no reason to believe that his constructed computational states are *even weakly* causally related; since Putnam can't appeal to the Principle of Noncyclical Behavior, he can't establish the disjointness of the s_i (cf. section 6.2).

There is another way (albeit one that requires much more elaboration than can be given here) that complexity considerations might tell against Putnam's argument. This is based on the insight that, roughly speaking, one's theory of a phenomenon should at least be *less complex* than the phenomenon itself. If it isn't, then the theory is in some sense confabulating, or at least not cutting nature at its joints. Suppose I present you with a steel ball, and claim that it is implementing a particular expert system, say Mycin. You ask me to substantiate this outrageous claim. I proceed to do so, by finding strange, relational, disjunctive, and complex characterizations of the steel ball states to identify with each of Mycin's computational states. This characterization would be so complex, in fact, that a text representation of it might take up, say, one thousand times the computer disk space that the Mycin program itself takes up! Anyway, I go on to claim that with this interpretation of the steel ball states, I can tell you how Mycin would respond to any given query. Even if I could, it would only be because of the complexity of the interpretation function, not the steel ball. The steel ball wouldn't be implementing Mycin, *I* would be. The intuition that this type of story is supposed to motivate is that it is natural to put some restrictions on the relative complexity of our interpretations in order to rule out such cases. Such restrictions would, no doubt, rule out Putnam's interpretations as well.[8]

Finally, it should be pointed out that computational descriptions do not only specify causal transitions that must take place; they also implicitly *prohibit* many transitions. For example, if an automaton is supposed to move causally from state A to state B, then it

[8] In thinking about the issues raised in the above passage, I benefited from a discussion with Matthew Elton.

is supposed to do this *without moving into state C in the process*. Putnam tries to avoid the difficulties that this observation raises by defining the s_i to be the region containing all of the states of S between t_i and $t_i + 1$. This would rule out the possibility of the S moving from A to B via C, but only if the s_i could be shown to be disjoint. But we have already seen that he cannot show this.

To summarize some of the main points so far, Putnam's argument for the universal realizability of finite automata is uncompelling because:

> The disjunctive nature of its individuation of computational states limits Putnam to post hoc *descriptive* states, yet computational characterizations are also *predictive*;
>
> Its notion of causation is too liberal, in that it would allow as causally related many events that, in everyday life and sciences other than mathematical physics, we would *not* take to be causally related;
>
> It relies on the Principle of Noncyclical Behavior and the lemma, which both, in turn, rely on an unconvincing and largely empirical account based on "clocks". Thus, it fails to establish that the transitions are even weakly causal, and fails to establish the disjointness of the realizing states;
>
> The failure to establish the disjointness of the realizing states yields ambiguous interpretation functions, and prevents Putnam from accounting for the fact that computational characterizations *prohibit* certain state transitions.

6.6 Computation and the world: Inputs and outputs

But wait; there's more. Computers don't, in general, *just* sit around making state transitions. They receive signals from keyboards, mice, and video cameras, and control displays, printers, and robot arms. They *do* things; they interact with things. Even formal automata include a notion of input and output. Another problem, then, for Putnam's proof is that, strictly speaking, he only establishes it for the case of automata without any inputs or outputs (Putnam admits as much on page 124). To try to rectify this, Putnam would have to count the state of the boundary of S at a particular time to be the input to, and output of, the automaton. Let $[A : I_i : O_j : B]$ indicate a finite automaton that when in state A, receives input I_i which causes it both to output O_j, and to move into state B. Putnam must define each I_i as the disjunction of all the boundaries of S that correspond to states which receive I_i as input, in the interval being interpreted. For example, consider the finite automaton: $[A : I_1 : O_1 : B][B : I_2 : O_2 : C][C : I_1 : O_1 : A]$. If physical state s_1 is interpreted as state A, s_3 is interpreted as C, then I_1 could be defined as: $boundary(s_1) \vee boundary(s_3)$. Only then can Putnam, in a way similar to before, argue that the computational state of the system and the input received in that state *jointly cause* the system to move into the next state, and emit an output.

One problem with this approach is that it isn't faithful to the notion of input and output that is involved in computation. For computational purposes, inputs and outputs are characterized in terms of their intrinsic properties. If we *define* inputs and outputs in a post hoc manner, as whatever boundary state a physical system has at a particular time, then Putnam's argument has a chance of going through.[9]

But if the definition of an output is fixed *in advance* as, say, the display of a character on a video display, then Putnam will not be able to show that a given system, for example my office wall, instantiates any formal automaton with that kind of output. That is because the state transitions of the wall will not causally determine the output, even, presumably, on Putnam's weak notion of causation. Varying the states of the wall (considering the various possible trajectories of the physical system with respect to its input) will not result in a corresponding variation in video display states. Therefore, the output is not *caused* by the state transitions in question. Similar considerations apply in the case of inputs. So only post hoc notions of input and output will allow Putnam to maintain his universal realizability thesis, yet post hoc notions are unacceptable for predictive and explanatory purposes. If what counts as a physical realization of an output is not fixed in advance, then we can guarantee that any system will emit a given output in the future, but only at the price of having no idea of how that output will be manifested. We will only have a descriptive, not a predictive computational understanding of the system (cf. the end of

[9] But even then one will be in the unsatisfying position of being unable to differentiate inputs from outputs, since they are both defined to be the same boundary state.

section 6.3).

In fact, Putnam admits that for any given automaton with inputs and outputs, one will be able to restrict the set of systems that instantiate it [Putnam, 1988, p. 124]. In some sense, then, he admits defeat: not *every* physical system can instantiate every finite automaton. But he doesn't really consider this concession to be a concession of defeat. That's because he believes that one will still have universal realizability of computation within the class of physical systems that get the input and output right:

> Imagine, however, that an object S which takes strings of "1"s as inputs and prints such strings as outputs behaves from 12:00 to 12:07 exactly as *if* it had a certain description D. That is, S receives a certain string, say "111111" at 12:00 and prints a certain string, say "11" at 12:07, and there "exists" (mathematically speaking) a machine with description D which does this (by being in the appropriate state at each of the specified intervals, say 12:00 to 12:01, 12:01 to 12:02,..., and printing or erasing what it is supposed to print or erase when it is in a given state and scanning a given symbol). In this case, S too can be *interpreted* as being in these same logical states $A, B, C, ...$ at the very same times and following the very same transition rules; that is to say, we can find *physical* states $A, B, C, ...$ which S possesses at the appropriate times and which stand in the appropriate causal relations to one another and to the inputs and outputs. The method of proof is exactly the same as in the theorem just proved (the unconstrained case). Thus we obtain that *the assumption that something is a "realization" of a given automaton description (possesses a specified "functional organization") is equivalent to the statement that it behaves as if it has that description* [Putnam, 1988, p. 124, his emphasis].

Putnam means "behaves" here purely externally: any physical system that, for a given time period, has the same inputs and outputs as a particular finite automaton, instantiates that automaton. Thus, Putnam is claiming that there is no computational difference between the two following systems:

> A program that calculates trajectories for spacecraft on the basis of certain input parameters (position, mass and velocity of the craft and nearby bodies) that is run, on three successive occasions, on the inputs a, b, c respectively and yields outputs x, y, z respectively;
>
> A lookup table which only has three entries: $a \mapsto x, b \mapsto y, c \mapsto z$.

Such an equivalence would be bad enough for our current understanding of computation, but Putnam has even more specific prey in mind. In particular, the reason why he is attempting to undermine computation in general is because he is opposed to its use as a foundation for an understanding of the mental in particular. And if Putnam can show that all behaviourally equivalent systems instantiate the same program, then he will have shown that functionalism implies behaviourism, a conclusion that many who wish to use computation as a foundation for cognition would be loathe to accept.

Of course, the conclusion need not be accepted, since it depends on the central argument of universal realizability, which, as we have seen, doesn't work. Nevertheless, one might think that the computational equivalence of behaviourally identical systems might have held *if* Putnam's original argument were sound. But I don't think even this is correct. Perhaps if one restricts oneself to characterizing a particular temporal interval of a system, then one could get the equivalence of behaviour and computation, if Putnam's main argument were successful. But this is to make the mistake (again) of seeing computational characterizations as purely descriptive, and not explanatory or predictive (cf. the end of section 6.3). Not all systems that have the same inputs and outputs for a short interval will continue to have the same inputs and outputs in the future. Thus, a particular computational characterization will apply, for predictive purposes, only to some small subset of those physical systems.

6.7 The worst case: universal, but useful

Input/output issues aside, one might think: OK, so Putnam doesn't show that every system realizes *every* finite automaton. There are, in principle, limits to what can count as an acceptable interpretation. But the fact is that, given the natural complexity of physical stuff out there, there is still a *lot* of room for indeterminacy. Even if every system doesn't instantiate *every* automaton, it might be that every ordinary macroscopic system (like a brain) instantiates an *infinite* number of automata.[10]

[10] Notice that the Cryptographer's Constraint, though useful in other contexts (viz. syntax to semantics, rather than physics to syntax considerations), doesn't help here. The Cryptographer's Constraint (which has been mentioned in related contexts by McCarthy, Dennett, and Harnad) is the observation that as, say, the length of a string of characters increases, the chances that there is more than one meaningful interpretation for that message decreases drastically. The reason why we cannot apply this constraint here (even assuming that we find some syntactic norm to replace the one of "meaningful") is that Putnam

As stated in section 6.1, such indeterminacy doesn't count against computation. There is a reason why Putnam set his goal to be such a lofty one: it is the only one which can really count against the ontological status of computation. It is only by guaranteeing that every system instantiates *every* program that one can be sure that no matter what computational account one gives of the brain, it will apply just as well to stones, roads, and walls. If it is admitted that some systems do not instantiate every program, then one will not be able to conclude that everything implements any particular program that cognitive science puts forward. That is, the modified claim allows computational characterizations to be non-vacuous, which in turn upholds the coherence of the computational approach to understanding the mind. Which is just what Putnam wishes to reject.

Thus, for computational states to be ontologically sound, one does not have to show that there is only *one*, *unique* computational characterization that applies to a given physical system. In fact, computational practice hinges on just the opposite: that a particular physical system can be understood to be instantiating simultaneously, say, a word-processing program, and a universal Turing Machine. That is, some degree of indeterminacy of computational description is acceptable, or even desirable.

But *what if* computation were universally realizable? What if, barring the just presented arguments to the contrary, any ordinary open physical system could be interpreted as, say, running any program? It is worthwhile to look at just what would follow from what Putnam is trying to establish.[11]

Even if everything is every kind of computer, the brute facts are: 1) we don't actually seek to understand everything in terms of computational properties; and 2) computational

is not allowing us (via his continuity assumption) to take as fixed in advance the primitives ("characters") over which the interpretation is being conducted. Consider: if a cryptographer doesn't even know what counts as the characters of a coded message (the prima facie characters? Their orthographical components? The tertiary structure of the molecules of ink?), then the Cryptographer's Constraint does not apply.

[11] To be fair, it should be pointed out, again, that Putnam's main goal in his text was to undermine any computational understanding of mind, and not necessarily anything more. Nevertheless, I sense that Putnam's general scepticism concerning the "reality" of computation is shared by an alarming number of people, many of whom apply it to a broader range of issues. Therefore, the further discussion here is relevant.

explanations, although limited, are actually satisfying in a large number of cases. This just shows that even if computationality is "merely attributed", it can nevertheless be explanatory.[12]

The fact is, it *is* very useful to understand many physical systems (IBM's, Suns, Macintoshes, etc.) in terms of computational properties; and there are many more systems for which such an understanding is *not* useful. If computational properties are universally realizable, this just shows that for some systems, we can always competently assign computational properties in such a way that such assignments will allow us to develop an explanatory and predictive understanding of those systems. If ontology is completely independent of these explanatory concerns, then perhaps claims of the form "physical system x instantiates automaton P" are meaningless in some absolute sense. But if explanatory (or even mere utility) considerations have any say in whether an attribution is warranted or not, then it is clear that sometimes we will be warranted in deeming a system a computer, sometimes not. The question "is this physical system a digital computer running program P?" will be meaningful, and resolved, at least to some degree, empirically.

6.8 Formal computation and relational explanation

Some may ask: why defend these formal models of computation, when there are many reasons to believe that more embedded, embodied and semantic accounts are required to understand real world computational systems? In particular, Peacocke has recently argued that formal computation cannot do the work many in cognitive science intend for it: to provide explanations of psychological events *qua* psychological events. If this is the case, then it was hardly worth defending formal computation against the criticisms of Searle and Putnam, correct though that defense may be, since formal computation will fail to be useful for the purposes at hand anyway. Therefore an examination of Peacocke's recent arguments against the explanatory power of formal computation [Peacocke, 1994a] is in order.

[12] In fact, there is a strong current in modern philosophy of science that claims that many, if not all of our explanatory sciences, even (or especially) those as fundamental as quantum physics, are based as much on human interests as they are on some ontologically independent reality.

Peacocke is concerned with a particular role for computation in a scientific psychology: providing explanations of psychological events. The first defense of formal computation, then, can be made immediately: there is at least one role for computation in a scientific psychology that is not of the form that Peacocke considers. Specifically, the role that has been emphasized so far in this work: naturalizing the intentional characterizations used in the the psychological explanations. For example, one might take a Fodorian view on scientific psychology, seeing it as completely free of computational notions, and being some refined version of folk psychology. Nevertheless, it might be of use to naturalize that intentional account using formal computational notions, in the familiar "syntax mirroring semantics" manner. But even if one did use computation of some sort in one's psychology, it would still remain possible for formal computation to play a naturalizing role as well. Since formal computation is not being asked to provide the explanations that Peacocke is considering, its failure to do so will not count against it.

This reply works only if the naturalization relation does not imply an explanation relation. Is it possible to naturalize an intentional characterization without *explaining* it in Peacocke's sense? It seems possible that one could show why it is not a miraculous coincidence [Cussins, 1987] that a system is correctly described both in particular intentional terms and in particular non-intentional terms, without explaining the occurrence of a psychological event *qua* psychological event. It seems possible that one could make intelligible the fact that a state with these non-intentional properties has these intentional properties (and vice versa) without being able to explain the occurrence of a particular psychological event that the intentional properties characterize. If so, then formal computation might still have a role to play in cognitive science.

In fact, it is arguable that the kind of intentionally-individuated computational states that Peacocke recommends for psychological explanation will themselves have to be naturalized. If so, then it could very well be that such a naturalization must proceed by appealing to non-intentionally individuated states, and by making it intelligible that these states are intentional states. And cognitive science is betting on the fact that in order to do the naturalizing work, the non-intentional characterization will have to be relatively abstract – that is to say, formal – as opposed to being, e.g, a biological or physical characterization.

Even if this is right, I think it is worth looking at the case Peacocke *does* have in mind, that of "explaining the occurrence of an event with relational properties."

One of Peacocke's arguments is, at root: psychological states are intentional; intentional states have relational properties which are partly constitutive of those states; explaining a state with relational properties *qua* a state with those intentional properties involves, at least, providing an explanation that supports the right counterfactuals. Formal computation *simpliciter* obviously cannot provide such an explanation, since the explanation proceeds quite independently of they way the external world is, and yet the way the external world is is exactly what one must take into account if one's explanation is to support counterfactuals that involve relational properties. This much seems uncontestable.

But Peacocke is quick to pour water on the fire of a different proposal, one which is an obvious extension of the purely formal mode of explanation just discounted. The proposal is that one can explain an event with relational properties by first explaining "the occurrence of an event with a property, intrinsically characterized in terms of bodily movements, and... add that it also stands in such and such relations." Let us call this the "dual component proposal". Peacocke's objection is that this proposal cannot be right, since if it were, "explaining an event under any of its descriptions would be explaining it under all of its descriptions – which must be overshooting."

But it is not clear that this undesirable consequence follows. The reason why one might think it does is because one might take the second of the two components to be a mere *stating* of relational properties, when instead it should really be taken to be an *explanation* or *naturalization* of the relational properties in terms of both the internal, formal states of the system and the state of the system's environment. Consider the case Peacocke mentions, that of explaining the occurrence of an interposition of one's hand between the sun and one's eyes. A dual component explanation would proceed by appealing to formal principles to explain a bodily movement, intrinsically characterized, and then appealing to external factors to show that *this* movement, in the context of *that* position of the sun, and *that* position of one's eyes, is consistent with the movement having another description, that of "being an interposition of one's hand between the sun and one's eyes." But more is required than mere consistency; it must also be the case that the explanation preserves the right counterfactuals, which it does. Had the sun been in a different place, then the input to the formal computation would have been different, thus the movement would have been different, and yet it still would have moved in such a way as to be consistent with being an interposition of one's hand between the sun and one's eyes.

Thus, the contribution of the external component is much more than a mere *statement* of the movement's relational properties; it is an explanation of why that intrinsic state plus that environmental configuration yields a potential instance of the relational properties under consideration. And it is this robustness which allows the dual component proposal to avoid Peacocke's objection. That explanation will *not* support other counterfactuals, counterfactuals relevant, say, to the movement being part of a signal in semaphore. The external component which explains why the movement constitutes a "shading" movement does not even *mention* the environmental components required to explain why it is part of a signal in semaphore, let alone relate them in the right ways. Therefore, the dual component explanation of the movement, under the description "an interposition of one's hand between the sun and one's eyes", does not also explain the movement under the description "a part of a signal in semaphore". Nor does it explain the movement under any other unrelated descriptions. So the overshoot is avoided.

Peacocke has another, more direct reason for thinking that formal computation cannot provide psychological explanations: the definition of "formal". He looks to Fodor to give the canonical account of formality:

> Formal operations are the ones that specified without reference to such semantic properties of representation as, for example, truth, reference and meaning [Fodor, 1981, p 309].
>
> If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the *semantic* properties of such representations, including... the property of being representations *of the environment* [Fodor, 1981, p 314, emphasis in original].

Peacocke is right to take this to be an uncontroversial view of formality, not unique to Fodor, but something quite consistent with denying the Language of Thought hypothesis. But in explaining this generality of Fodor's notion of formality, Peacocke reveals that he is reading much more into these definitions than is strictly warranted:

> It would prima facie be entirely consistent... to hold simultaneously all these propositions: that some intentional states are realized in non-sentential states of a connectionist network; that some sequences of states of the network are computations; and that semantic properties are never involved in either the explaining or the explained states of a computational explanation [Peacocke, 1994a].

It seems an illicit move has been made from understanding formal computational states as not involving semantics in their *specification*, to understanding them as not involving semantics in the *explanations* they provide. This is repeated later:

> We can elaborate a little what the mistake of principle would be by thinking about a putative computational explanation of a person's coming to be in an intentional state. On the non-semantic view of computation, this involves one non-semantic state explaining, by some computational procedure, a second computational state. This second state is said to be the basis of (or realization of, or what constitutes) the intentional state to be explained. But if only non-semantic properties are explained, where is the explanation of the intentional properties? It seems that on the non-semantic conception of computation, only non-semantic features of intentional states could be explained [Peacocke, 1994a].

Surely, if formal computational states were *defined* to be states which cannot explain semantic states, then their role in cognitive science would be bleak from the start. But the point that seems to be missed is this: that a state is *specified* in a non-semantical way does not in itself imply that it *is* non-semantic, nor does it in itself imply that the state can only *explain* non-semantic features of intentional states.[13]

Peacocke argues that including the environment will not help here:

> If the explaining conditions of the computational explanation were externally individuated, there might still be some room for manoeuvre here – perhaps the explaining conditions could ensure the right external relations required for the intentional state to have the content it does. But this too is ruled out by the non-semantic conception of computation, according to which the explaining conditions are non-semantic too. On that conception, the internally individuated explaining conditions cannot magic into existence the complex of external relations required for an intentional state.

[13]To be fair, I think the second of Fodor's two characterizations of formality, above, is misleading, and invites misunderstanding in this regard. That is, to keep in spirit with the first characterization, Fodor should not be saying that formal process do not in fact have access to semantic properties; rather, he should only be claiming that formal processes may be individuated and understood without reference to the access of any semantic properties. But even if one (perversely) insisted on using the letter, and not the spirit, of Fodor's second characterization, that would still only imply that there is a *metaphysical* gap between formal and semantic states, not an *explanatory* one.

He then goes on to say as much for the presupposed background conditions of computational explanations: since such explanations are non-semantic, so must also be their background conditions.

Once again, it seems that these apparent consequences of the definition of formality are illusory. There is, in the characterizations of formality given, no mention of explanation at all, let alone restrictions on explaining or background conditions. In fact, the ever popular dual component approach to giving computational explanations of intentional properties, discussed above, explicitly violates this preclusion of externally individuated explaining conditions. But even if one did have some reason for thinking that formality should disallow *semantically* individuated explaining conditions, this would not mean that *externally* individuated conditions would be thereby prohibited. In the example of a dual component explanation given before (shielding one's eyes), externally individuated explaining conditions were used, not semantically individuated ones. Yet it was argued that there is no reason to deny that these conditions, together with an account involving formal states, could explain the occurrence of an intentional event *qua* intentional event.

This defense of formality should *not* be construed as the claim that computation is formal. I think it is unfortunate that these formal, abstract automata are thought of as capturing the heart of computation. I hope I have made it clear that they do not; computation is essentially semantic, embedded and embodied, none of which properties are captured by formal automata. Nor am I denying the external nature of intentional states, and the requirements for externally individuated properties in their explanation. It is just that formal properties have a role, perhaps a necessary one, both in furthering our understanding of computation, and in providing naturalistic explanations of intentionality. Or at least, I am not yet convinced that they have no such role.

6.9 Formal computation: meaningful, but inadequate

To be fair, characterizing computation in terms of actual inputs and outputs, and in such a way that the actual causal properties of the underlying physical system matter, ventures far beyond the explicit nature of current computational theory, as expressed in, for example, Turing Machines and recursive function theory.[14]

In fact, some may ask: why defend these formal models of computation, when there are many reasons to believe that more embedded, embodied and semantic accounts are required to understand real world computational systems? I agree that a theory of computation founded solely upon formal notions such as Turing Machines and finite state automata would be an impoverished one. Nevertheless, I think that it would be premature to assume that the success of a mature theory of computation is independent of the status of these purely formal theories.

Accordingly, both Putnam and Searle have done cognitive science a service, by drawing attention to the fact that its uses of the notion of computation may only make sense when accompanied by some implicit assumptions. These assumptions should be made explicit, so that they may be developed and refined. Both Searle and Putnam are in one sense right: a completely formal, non-causal notion of computation is inappropriate for cognitive science. Fortunately, our current understanding, at least implicitly, is more concrete: it is not empty and incoherent (as they claim). Nevertheless, those of us who wish to understand computation, especially those who wish to understand how it relates to cognition, have a substantial and exciting task ahead: that of discovering and articulating these non-formal elements of computation, whether they are, like causation and embeddedness, implicit in our current understanding, or as yet unknown.

[14] However: 1) some theorists are trying to correct this, as Searle points out [Searle, 1992, p 209]; see, e.g., [Smith, 1992] and [Smith, 1996]; 2) although embedded, causal computation might be at odds with our current theoretical understanding of computation, it doesn't seem to be that alien to our everyday notion of computation as manifested in computational *practice*.

CHAPTER 7

Sub-symbolic naturalizations of non-conceptual content

7.1 Content & vehicle: The NCC/PDP connection

In the cognitive scientific explanatory scheme, contents need vehicles. Naturalism requires that we show how our intentional and non-intentional characterizations of an organism cohere; the hypothesis of cognitive science is that we should do this by finding a computational characterization (vehicle) which marches in step with our non-conceptual intentional account (content) [Cussins, 1987, Fodor, 1985]. There is reason to believe that such a computational architecture will differ substantially from classical, purely symbolic notions of computation, which are typically meant to correspond to objective, conceptual contents.

There are those few who not only maintain that there is a significant difference between symbolic representation and the kind of representation one finds in parallel distributed processing (PDP)[1] models of cognition, but who also connect this difference with a difference in the kind of content carried by each of these forms of representation (e.g., [Haugeland, 1991], [Cussins, 1990]); I count myself as among this number. The general idea is that whereas symbolic representations carry conceptual, systematic content, composed of elements corresponding to objects and properties, PDP representations carry non-conceptual, non-systematic content that is composed of elements that do not correspond to orthodox objects and properties. This chapter discusses why one might think parallel distributed processing and non-conceptual content are a natural match. The affinities are

[1] For terminology buffs: I continue to use the designation "Parallel Distributed Processing" or "PDP" instead of, e.g., the less passé "connectionism" because it gives a better indication of what is central to the relevance of such networks to this research.

made precise by citing examples involving a particular connectionist cognitive mapping network: the Connectionist Navigational Map.

In a way, some authors (e.g., [Fodor and Pylyshyn, 1988], [Davies, 1990], [Ramsey et al., 1991]), although they draw some conclusions that are radically opposed to the ones proposed here, have done a lot of this chapter's work for me. Their arguments against the compatibility of PDP and the propositional attitudes implicitly assume that all content is conceptual content. "[Conceptual] content", the arguments go, "can only be carried by representational vehicles that do not lack property X[2]. But PDP vehicles lack property X. Therefore, PDP representational vehicles cannot carry [conceptual] content." Rather than concluding from these arguments that PDP is uninteresting and irrelevant to cognitive science, one can instead conclude that these arguments establish PDP (and NCC) as a radical alternative to the orthodox: if PDP representations do not carry systematic, conceptual content, then (since they evidently carry *some* kind of content) they must carry some other, non-systematic kind of content. *Non*-conceptual content.

From the point of view being developed here, then, PDP and NCC need each other:

> PDP needs NCC. If there is no NCC, and if the arguments that attempt to show the incompatibility of PDP and conceptual content are right, then, despite appearances to the contrary, PDP representations carry no (explicable) content at all. Thus, PDP cannot be a cognitive architecture: it cannot naturalize intentional characterizations.
>
> NCC needs PDP. As stated before, naturalism demands that a characterization of an organism at the level of content, which we cannot directly understand to be manifested by a non-intentional characterization of that organism, be shown to "march in step" with a characterization (in cognitive science, this characterization is computational) of that organism that we *can* understand to be so manifested. One of the principal attractions of Fodor's Language of Thought approach to cognition is the Representational Theory of Mind which underlies it [Fodor, 1987]. The Representational Theory of Mind attempts to explain how intentional phenomena could

[2]Where X is, variously: necessary systematicity; structured representations and operations sensitive to that structure; isolable, causally efficacious syntactic entities that mirror conceptual contents; or whatever.

> be realized in the physical world by showing how the physical (characterized in terms of classical computational states) and the intentional (characterized in terms of conceptual semantics) march in step. To be naturalized, then, NCC will require a computational architecture with which it can march in step. Perhaps PDP is such an architecture; there might be others (in which case, Cussins [Cussins, 1990, page 431] is right: PDP needs NCC more than the converse).

However, it is not wise to rely too heavily on these arguments that others have provided, as they establish too much: that PDP could *never* carry conceptual content. Since I want to explore the possibility that PDP could provide the architecture for an explanation of the development from NCC to objective, conceptual representation, the full conclusions of these arguments should be avoided if possible. My position is that, unlike classical cognitive architectures, PDP architectures can be appropriate for both non-conceptual *and* conceptual contents – depending on the stage of development of the architecture. I thus make a distinction between architectures that are *necessarily* systematic, and ones for which particular configurations result in *contingent* systematicity. The basic insight is that there are non-systematic intentional phenomena, so cognitive science should favour the latter type of architecture, which can accommodate both types of intentionality.

Even if this position is taken, there remains an alternative, related argument against PDP as a cognitive architecture to be found in Fodor and Pylyshyn's paper. This argument does *not* assume that systematic, conceptual contents can be carried only by representational vehicles that are *necessarily* systematic, but allows that *contingently* systematic vehicles can do the trick. The argument takes the form of a dilemma: either PDP can attain the (contingent) systematicity required for conceptual content, or it cannot:

> Horn 1: if PDP *can* achieve systematicity, then it is uninteresting to cognitive science; it is a mere implementation of the classical architecture (structured representations and operations sensitive to that structure);
>
> Horn 2: if PDP *cannot* achieve systematicity, then it cannot model human competence, such as the productivity of language (e.g., the ability to understand arbitrary novel sentences).

Some might wish to reject Horn 2, and deny that human behaviour is ever such that it must be understood to be the product of systematic mental processes; I need not make such a bold claim here. Rather, I concede that PDP's success as an architecture for *all* of human cognition requires that it be able to construct systematic representations: PDP must be able to *attain* (contingent) systematicity.[3] Otherwise, we would have an unattractive, inexplicable schism in our understanding of cognition: the (necessarily) non-systematic, PDP-modelled phenomena, incommensurable with the (necessarily) systematic, classically-modelled phenomena. But the best theory will be the one that does *not* leave the transition from non-systematic subjectivity to systematic objectivity a mystery, but explains it. The existence of developmental data provides an extra, crucial constraint on our cognitive architectures. The point being made is not just that one should prefer architectures whose existence can be explained to those which are miraculous; one should also prefer, ceteris paribus, architectures for which there is a *better* explanation of its development to those with a *worse* explanation. The advantage of a non-conceptual approach to cognitive architecture is that the explanation of systematicity *is itself intentional*; in a classical system, such an explanation, if there is one, must remain non-intentional and mechanistic, with all the attendant disadvantages in terms of explanatory power, generalizability, etc.

Thus (to get back to the dilemma), Horn 2 cannot be denied; it is Horn 1 which should be resisted. The supporter of PDP in general could do this in a number of ways (although only the first response is appropriate here):

1. First and foremost: agreement in a special or limiting case is not the same as implementation; a PDP architecture that meets the classical constraints in the special case of systematic adult human behaviour is no more irrelevant to cognitive sci-

[3] And there is work [Chalmers, 1990, Elman, 1990, Pollack, 1990, van Gelder, 1990] that *suggests* that PDP *can* achieve such contingent systematicity. Of course, it is clear that there could be a PDP implementation of a systematic architecture. But this would be as insufficient for cognitive science (in the broad, Recalcitrant-Phenomena-involving sense of "cognitive science" being used in this chapter) as any other classical architecture, and for the same reasons. That is, it would not be able to account for the non-systematic phenomena, and it would leave inexplicable the transitions between non-systematicity and systematicity – the development of objectivity.

ence than a quantum theory that meets Newtonian constraints in the special case of macroscopic conditions is irrelevant to physics (after Smolensky [Smolensky, 1988]);

2. Even if PDP's contingent systematicity *did* render it a strict implementation of classical architectures, that would still allow it to facilitate the construction of theories with explanations and predictions considerably different from theories built on other classical architectures. To be a member of the class of classical architectures, one need only meet two very general constraints: possess structured representations and structure-sensitive operations. Thus, theories based on PDP architectures would be just as relevant to cognitive science as those based on architectures like SOAR [Rosenbloom et al., 1992] or ACT* [Anderson, 1983], which also respect the classical constraints. There is no reason to believe that just because two architectures meet the classical constraints that all the theories specifiable in terms of one architecture will also be specifiable in terms of the other. If PDP is mere implementation theory because it meets the classical constraints, so are SOAR and ACT*. But surely Fodor and Pylyshyn don't want to claim that SOAR and ACT* are therefore irrelevant to cognitive science;

3. Also, "mere" implementation theory isn't that irrelevant to cognitive science. As said before, naturalism demands that the content-involving intentional level be shown to "march in step" with the computational architecture. A computational architecture which met this demand would be of reduced value if it could not be shown to cohere with a non-computational, physiological characterization of the organism whose psychology we were striving to understand. In this respect, not all implementations are created equal: we should prefer the one that can be understood to be instantiated by the organism;

4. Last (though least likely), PDP *might* be able to achieve systematicity (model our generative, productive, etc., behaviour) other than by meeting the classical constraints, *pace* Fodor and Pylyshyn's claim that theirs is the only game in town.

In summary, Fodor and Pylyshyn were right to admit that PDP might be able to achieve contingent systematicity, but they were mistaken in failing to realize that such a characteristic is not a detraction, but a desideratum.

7.2 A case study of PDP/NCC affinities

With the foregoing by way of introducing the notion of non-conceptual content and the theoretical issues concerning its connection with parallel distributed processing, the remainder of this chapter employs the development of a particular PDP network in order to illustrate the fact that the contents of at least some PDP representations are best viewed as non-conceptual. This is done by giving five examples of how such contents cannot be understood to be conceptual. The next chapter demonstrates the transition from representations with less systematic contents to those with more systematic contents: the development of conceptuality. But first, the network architecture itself is described.

7.3 The Connectionist navigational map

This section reviews the CNM architecture, environment, and learning regime.

7.3.1 CNM architecture

The *Connectionist Navigational Map* is a computational architecture being developed with the aim of providing an autonomous robot with the ability to learn and use spatial maps for navigation. One component of this architecture, the *predictive map*, allows the robot to predict what sensations it would have if it were to move in a particular ego-centrically specified manner (e.g "rotate $\pi/4$ radians to the right", "move forward 10 feet"). Of course, this requires the robot to have some kind of representation of its current location, since, in general, the mapping from actions to sensations is dependent upon where one is in the world. That is, the mapping from sensations and actions to sensations is one-to-many, since more than one place can have any given sensory signature. Thus, the spatial environment, and therefore a model of it, can be seen instead as a function from current location and current action to predicted sensations. The input consists of a state representation, or *location code*, corresponding to the current location a of the robot, and an action representation representing the move m being made. The output of the network

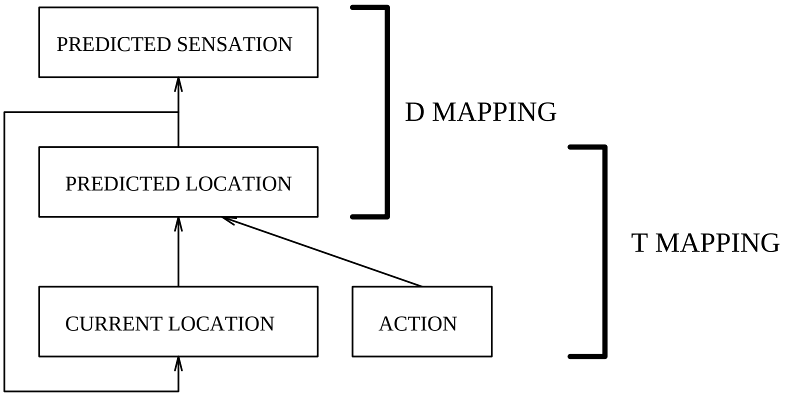

Figure 7.1: The PDP architecture of the predictive map.

is a vector that is supposed to be equal to the sensation vector the robot would receive from its senses if it were actually at the place that is reached by making the move m at location a.

Of course, there is more structure to space than a simple, direct mapping from locations and actions to sensations indicates. Specifically, location and action determine a new location, which itself determines the sensations of the robot. Thus, it might be easier for a robot to learn (or a theorist to analyse) a predictive map if its structure reflects this regularity of the spatial environment. The predictive map of the CNM is thus a composition of two mappings: a topological mapping T (from locations and actions to states) and a descriptive mapping D (from locations to sensations). In actual use, the location output of the T mapping, after a given action, is used as the location input to the T mapping for the next action. This is shown in figure 7.1 (arrows indicate directed, full inter-connection between layers of units).

Thus, if a constantly north-facing robot considers moving forward and then moving right, it can use the map to predict what sensations it would have after those moves by calculating $D(T(T(a,\textbf{move-north}), \textbf{move-east}))$, where a is a location representation corresponding to the robot's initial location before the actions, and **move-north** and **move-east** are action representations with the intuitive interpretation.

The predictive map can be realized in a hybrid system, with the topological mapping realized symbolically, and the descriptive mapping realized by a PDP network. It is argued in [Chrisley, 1991] that this kind of hybrid structure is attractive for the purposes of engineering a working system; *however*, if the primary motivation for constructing the system is to understand pre-objective representations, and how a robot can make the the transition from pre-objective to objective representations of space (as it is in this and the next chapter), then a uniform, sub-symbolic architecture for both mappings should be employed.

Given the iterative nature of the T mapping, the predictive map must be a recurrent network; in the experiments discussed here, it is implemented as a *simple* recurrent network [Elman, 1990].

7.3.2 The experimental setup

Although the CNM is intended for actual robots moving in real space, it can be of use here in making concrete some of the connections between NCC and its computational vehicles. The experimental situation used here is a deliberately impoverished one: a developing (learning) agent moving through a simulated "grid world"; the part of the world simulated has only 81 cells or locations (9 by 9). Each location has a 4-bit vector associated with it, which can be understood to be the sensations the agent has when at that location. This can be seen in figure 7.2, which shows the region of the "grid world" used in the simulations. Only the details for the locations on the used route are shown. Arrows indicate the direction of travel along the route. The four-bit binary vector at each location indicates the description or sensation vector associated with that location.

As in its normal use, the CNM is to provide a means for this agent to improve its navigation of its space (and thus increase the objectivity of its contents concerning that space) through sensory prediction. This situation is grossly impoverished with respect to real-world cognitive map building, but the simplifications are not only acceptable, but necessary, given the nature of this discussion: *explication* of the affinities between NCC and PDP.

The agent has eight actions available at any location, those of moving into each of the adjacent locations (orientation is not modelled: the agent can be thought of as always facing north).

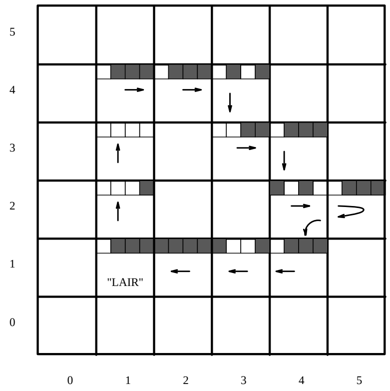

Figure 7.2: The region of the "grid world" used in the simulations.

It is assumed that the developing agent has somehow managed to conduct a journey that starts and ends at some privileged location, called *home*.[4] The developing agent stores the sequence of actions taken and the sensations that result. Note that sometimes the same action yields different sensations, and that different actions sometimes result in the same sensations.

7.3.3 Learning regime

When the agent returns home, it iteratively learns the route, not by actually moving, but by reviewing the remembered route in the following manner:

First, generate a training set:

1. Assume some arbitrary representation or code for the initial location ("home"). Store this code, and the code for the first action taken, as an input pattern; store the sensations that were observed after taking that action as the target output for that pattern.
2. Propagate the current input pattern into the T mapping.
3. Use the output of the T mapping, along with the next action on the remembered route, as the next current input pattern. Store this pattern into the training

[4]This return journey may have been the result of the developing agent being escorted by a parent, in which case, the cognition under investigation here is at least partly social, in that it relies upon the cooperation of multiple agents, and the interaction between agents and the social technology of *routes*.

list as an input pattern, as well as storing the next remembered sensation as the target output pattern for that input.

4. Go to 2 until finished with the remembered route.

Then, learn with the current training set: adjust the weights of the T and D mappings according to the gradient of the error (difference between target and actual outputs); i.e., use the back-propagation learning algorithm [Rumelhart et al., 1986].

After some period of learning with one training set (in the simulations described, the time was 6 to 10 epochs), a new training set is created, in the same manner as before (steps 1-4), and the agent trains with the new set.

In the simulation used for the experiments which follow, this 6 epoch cycle was repeated until the network had learned the route. That is, starting with the code for home and the initial action taken, the T mapping would produce a new code that not only yielded the correct predicted sensations via the D mapping, but which also, in conjunction with the representation for the next action taken, produced a code via the T mapping which could itself yield both the right sensation vector and location code, and so on, iteratively.

The T mapping was realized by a network which comprised 8 or 9 inputs (4 or 5 for the location code, and 4 for the action code; action codes were the 4 unit vectors for north, south, east and west, or the sum of the relevant unit vectors for the other 4 directions: NE, SE, SW and NW), 4 or 6 hidden units, and 4 or 5 outputs (a location code). The D mapping was implemented by a network with 5 inputs for a location code (in practice, this was the same as the output layer of the T mapping), and 4 outputs for a sensation vector at that location. For some simulations, a modified version of Fahlman's "Quickprop" algorithm [Fahlman, 1988] was used for speeding up learning. In all cases, the network was simply recurrent; back-propagation through time was not necessary.

In analysing the representational contents of this network, the activity patterns of the location, action, and sensation vectors were the vehicles chosen for consideration. This bypasses several very important issues (Don't weights serve as representational vehicles in the CNM? Shouldn't the representational vehicles include aspects of the network's environment?, et al.) which cannot be considered here.

7.4 CNM representations cannot be analysed conceptually (examples 1-5):

Of course, most things cannot be analysed conceptually. But that is because a stone cannot be analysed as carrying *any* kind of content. The examples that follow are meant to show why a particular PDP network that is, unlike a stone, intuitively a representational, content-involving system, cannot be understood conceptually, but is best analysed as containing representations with non-conceptual content.

7.4.1 Example 1: Perspective-dependent representations

In chapter 3 I argued for a connection between conceptuality and a form of objectivity. A good way to see the connection between one kind of systematicity and objectivity is to examine the limitations of the CNM T mapping. It is a consequence of the non-systematicity of T mappings that are learned just for the sake of navigating a few routes that their place representations will not be objective.

Forget, for the moment, about the particular route mentioned above, and consider the nature of T mappings in general. For example, suppose that the T mapping is non-systematic in that not all representations are mapped to four new, distinct representations. Suppose that only a few (enough to provide competence on the particular routes travelled) locations around the lair are systematically represented in the T mapping. This situation is illustrated in figure 7.3. Location codes are represented by circles: e.g. the circle marked "L" denotes the location code L which the agent is using to represent the lair; the code $T(L,$**move-east**$)$ is denoted by a circle that is pointed to by an arrow from the east (that is, right) side of the circle marked "L", etc.

It might be that this non-systematic topology is a *direct product* (albeit sub-optimal one) of the CNM's attempt to account for the data it has already encountered; perhaps there is something about what is in the space, like a pit at the place ten units east of the lair, which makes this sub-optimal, non-systematic topology quite practical in that context. Or it could be that the non-systematicity in regions not close to the lair is a mere *by-product* of developing a systematic T mapping for the area near the lair (perhaps like flattening a lump in a carpet merely moves the lump to some other part of the

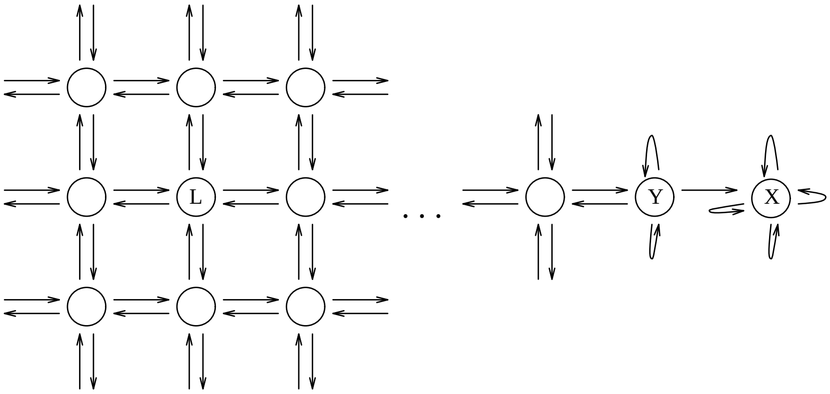

Figure 7.3: An example of local systematicity but global non-systematicity in the T mapping.

carpet). The reasons for the lack of systematicity are at worst irrelevant to, and at best supportive of, the purposes of this chapter (i.e., if it turns out to be practical to have a non-systematic topology in some contexts, this gives non-conceptual content an even more central role in psychological explanation, as a cognitive virtue in itself, as opposed to being a mere transitional state between oblivion and conceptuality, or between different conceptualizations). What is important is that we have a case where there is enough systematicity to get a notion of representational content going, but not enough to be able to speak of full Generality and systematicity.

In order to see why this non-systematic representation of space is non-objective, we have to notice a few constraints on objective space in general:

1. If one moves (in a very abstract sense of "moves") in a straight line from a location a, one will be located at a location $b \neq a$;

2. Moving in distinct directions from the same place will take one to distinct locations;

3. For every simple movement m there is an inverse m^{-1} such that $T(T(a,m),m^{-1}) = a$.

The T mapping illustrated in figure 7.3 is incapable of representing a space that meets these constraints. For example, once one iterates enough times through the T mapping, to calculate the representation for the location one gets to by, say, starting at the lair and executing **move-east** ten times, the resulting location code, X (represented by the circle marked "X" in the diagram), is such that:

1. $T(X,\textbf{move-east}) = X$, which violates constraint 1;

2. $T(X,\textbf{move-east}) = Y = T(X,\textbf{move-south})$, which violates constraint 2;

3. $T(Y,\textbf{move-east}) = X$, and $\textbf{move-west} = \textbf{move-east}^{-1}$, yet $T(X,\textbf{move-west}) \neq Y$, which violates the third constraint.

In such a case, the T mapping breaks down to such an extent that one cannot interpret the resulting location code, X, as a representation of the place ten units east from the lair.[5] Thus, the network cannot, as things stand, represent the location in question *at all.* This alone is sufficient to establish that, in such a case, even the location codes that *can* be understood to represent locations (i.e., those codes near the lair code, L) do not represent objective, conceptual locations, since truly objective locations are part of an arbitrarily extendible space, while these locations are not. These codes, unlike X, *can* support a sizeable, though limited, region of intentional spatial activity; for these codes, unlike for X, there is enough systematicity to determine a referent. But there isn't enough systematicity for that referent to be represented *as* an objective, perspective-independent place, a part of a purely objective space. Thus, the location codes, even those near (and including) L, carry non-conceptual content. Similarly for the action vectors: truly conceptual, objective movement should be arbitrarily iterative.

The fact that the violation of these constraints implies a non-objective representation of space can be shown in the following. The only way an agent with such a CNM configuration *could* represent (at this stage of development of its T mapping) the place ten units east would be by actually *moving* to that location. Upon moving east ten times, its location code for its current location would, by hypothesis, be of no use (e.g., it would not allow any successful predictions of the result of any action the agent might take). Keeping the T mapping fixed, the agent would have to co-opt the T mapping for its new locale, by

[5] Furthermore, one will not be able to understand it as representing *any* objective place, nor will one be able to understand *any other* location code as representing the place ten units east from the lair.

using L to represent not the lair, but the current location, and altering the D mapping accordingly. The location codes would have to undergo a change of referential significance (of a kind that would be impossible within the context-invariant restrictions of a classical architecture; see example 3 in section 7.4.3, below). But this means (assuming that the non-systematicity of the T mapping crops up when iterating **move-west** as well) that the agent will no longer be able to represent the space around the old lair; *after* the change, the lair will be just as unrepresentable, mutatis mutandis, as the current location was *before* moving.

Thus, the agent's ability to represent a place is dependent on the agent's *perspective*: in this case, its actual proximity to that place. Like the infant without the concept of object permanence, the agent would lack the notion of *location* permanence, to some *variable* extent. The CNM can represent places (i.e., it doesn't treat all qualitatively identical places as the same place, but individuates locations using some notion of non-locally-observable, relational properties). But since *objective* places do not disappear when one moves away from them, the network cannot be representing places *as* objective, perspective-independent places. Thus, the contents of the location codes of the network at such a stage in its development must be pre-objective, non-conceptual.

Another, brief way of putting the point is this: the relations between objective places are independent of both what is located at those places, and of our abilities to move between those places, whereas a non-systematic T mapping typically is not independent of these.[6]

[6] One might think, then, that the way to increase one's objectivity is to make one's T mapping of a region more and more independent of what one believes to be located in that region. This is to equate objectivity with Generality and systematicity, as I have been doing so far. But there is another view of what objectivity is: the ability arbitrarily to change one's T mapping, depending on, e.g., whether one is thinking of the movement of oneself, air, or light. That is, perhaps objectivity isn't Evans' "thinking from no point of view"[Evans, 1982, page 152], nor Nagel's "view from nowhere"[Nagel, 1986], but rather Smith's "view from somewhere"[Smith, 1992], or, perhaps better, "a view from anywhere"[Cussins, 1990, note 103, page 428]. This latter view of objectivity would seem to give even more pride of place to non-systematic architectures such as PDP.

7.4.2 Example 2: Non-systematicity

In some absolute sense, the above example of perspective-dependent representations already demonstrated the non-systematicity of such representations. But it demonstrates an asymmetry between the places that it can represent and those it cannot, whereas the Generality Constraint (cf. chapter 3) was meant only to impose systematicity among contents that the agent can actually entertain. Nevertheless, the CNM also exhibits this kind of non-systematicity among the places it *can* represent.

Perhaps the easiest way of seeing this is to consider the fact that the network's choice of codes to use in the T mapping will constrain the possibilities for the D mapping. For example, suppose that the T mapping has very similar codes for two places. Now one of the main features of PDP representation is that all of the representation vectors exist within a common relational space. The ability for PDP networks to generalize in an interpolating (and extrapolating) fashion stems from the fact that the requirement of producing a particular output vector for a given input vector will severely constrain the possible output vectors for any nearby input vector. Thus, networks cannot, in general, assign arbitrarily dissimilar outputs to arbitrarily similar inputs. This means it will be difficult for the D mapping to assign different sensory properties to similar location codes. There will be some predicational combinations (of sensory properties to locations) which will be difficult or impossible, violating the Generality Constraint. Thus, PDP's generalization ability is, in this sense, inherently non-systematic, as it violates the independence of subject and predicate.

7.4.3 Example 3: Context-sensitivity I

A third way of motivating the link between PDP and NCC focuses on the methodology of content ascription in both symbolic and PDP architectures. In symbolic architectures, the compositional nature of the semantics encourages theorists to make their primary semantic ascriptions on the atomic level. They assign (conceptual) contents to the atomic symbols; familiar recursive rules then determine the content of any molecular representation from the contents of its atomic constituents, together with their mode of combination. So the symbols **block-57** and **block-23** are assigned particular blocks as referents, **is-above** is assigned the "above" relation, and **is-above(block-57, block–23)** is assigned the value *true* if the first block is above the second block, and *false* otherwise. The procedure is to assign a semantic value to the atomic symbols[7], and then, given the compositional rules which determine the semantics of complex expressions, try to build a system that uses these symbols so that the conceptual ascriptions, to both the atomic and molecular symbols, are warranted. The compositional relations are built-in and guaranteed; what has to be achieved is a complex function of operation, both internal and interactive with the environment, that bestows on the atomic symbols the content optimistically assigned to them.

In contrast, the primary locus of PDP content ascriptions is the whole content level. For example, consider the following system. An autonomous robot receives sonar input from the 360° ring in its two-dimensional plane of action. Its perception and action are coordinated by a PDP network which has learned, not only to associate different sonar signatures with obstacles, but also to associate with the input an appropriate avoidance behaviour depending on both the type of signature (person vs. chair) and the position of the obstacle signature in the input ring (immediately to the left, straight ahead ten feet or so, etc.). In giving an intentional analysis of this network, a theorist might note that a certain pattern in the hidden units is always present when the network classifies the input as *chair that is in front and within three feet*. The hidden unit pattern in question is assigned the content *there's a chair less than three feet in front* or something similar. But there is no guarantee that there is any isolable part of this hidden unit representation that corresponds, in all situations, to the concept *chair* or the relation *being less than three feet in front*. For example, a hidden unit vector with the content *there's a chair over ten feet behind* might be completely *orthogonal* to the one about the chair being close and in front. The situation is the converse of the one for typical means of content ascription to symbolic systems: the content ascription is warranted, but the structure of the representation is in question. Thus, even if a hidden unit pattern in a more advanced PDP system were ascribed a content that involved a *particular* object, such as *the chair bumped into two*

[7]This is typically done with a large degree of wishful thinking; the comments in [McDermott, 1981] continue to be germane.

minutes ago is now behind, there would be few, if any, constraints on what hidden unit pattern is carrying the content concerning the chair that the robot bumped into two minutes ago. In fact, it might be that there is *no* context-invariant part of the hidden unit pattern that corresponds to the chair at all. The object/property decomposition is not built-in.

It is this difference in content ascription, I believe, that is responsible for two different kinds of context-sensitivity in PDP representations that makes PDP especially suited for NCC. I visualize these two kinds of context-sensitivity as functions from (sub-total) representations to (partial) contents[8], and call them *one-to-many* and *many-to-one* context-sensitivity, respectively.

The first kind of context-sensitivity is manifested in the fact that the same (sub-total) representation will carry many different contents, depending on the context. As said before, the fact that the locus of the primary content ascriptions for PDP representations is at the whole content level means that there is no built-in semantic relationship between the representations of whole contents and those of partial contents, and therefore there is nothing to prevent the same sub-total representation from being a constituent in two quite different representations, even two that have no overlapping content: *the chair is ahead* and *the person is behind*, say. In such a case, the content of the sub-total representation must be different in the two contexts, since there is no partial content common to both contexts that could be assigned to the representation. Note that this context-sensitivity is ruled out for symbolic representations, since the atoms are typically assigned context-invariant contents, and the contents of molecular symbols are determined from these atoms in a context-independent manner.

In our case of an agent learning with the CNM, this kind of context-sensitivity could arise if the same code were used for two different places simultaneously. That is, a case where, say, there are two regions of the the space that are sufficiently similar that the CNM

[8] By sub-total representations I mean sets of hidden units that are strictly less than the set of all hidden units in the system. All I mean to achieve by using this term is an emphasis that I am not talking about representations with whole contents (which in simple systems might not vary their contents across contexts), but rather constituent sub-patterns of such representations. For the whole content/partial content distinction, see chapter 2.

can use the same representation for two places, one in each of the similar regions. This will result in a lack of systematicity. If two place representations are not co-referential, then they must have different content (since content determines reference). If these contents were conceptual, and met the Generality Constraint, then it should be possible to arbitrarily combine these contents with predicational contents; it should be possible for the network to have a D mapping for one of the representations that is different from the D mapping for the other. But since the representations are identical in this case, there is no such possibility. The contents of the representations are not independent of each other, so they are not representing the places as conceptual, objective places; they must be representing them non-conceptually.

In the context of other networks, this one-to-many type of context sensitivity can imply non-systematicity in a different way. Since the content of a given sub-total representation depends on the context of the whole content representation in which it finds itself, it is possible that the sub-total representation might warrant the assignment of a particular content (e.g. *is green*) only in the context of whole contents involving food, say, and not predators (to use a slightly more biological example). Thus, the way that the system is representing *is green* does not meet the Generality Constraint, since the system cannot represent the whole content *the predator is green* even though it can represent other whole contents involving greenness and predators. Thus, the content of the representation in question must be non-conceptual. Not that we didn't already suspect, or even know, that PDP representations are not necessarily systematic; but this provides another way of seeing why non-guaranteed systematicity, and therefore NCC, is so fundamental to PDP representation: it is bound up with the means of content ascription used for them.

7.4.4 Example 4: Context sensitivity II

It's a familiar point that very often, one has to change oneself in order to maintain a stable link with a particular aspect of a changing environment. If one wants to continue to point to an airplane that is moving across the sky, then one will have to move one's arm, head, and torso. This basic idea underlies the second, many-to-one kind of context-sensitivity: several different sub-total representations being assigned roughly the same (conceptual) content. This is the phenomenon illustrated by the now-familiar coffee example described by Smolensky [Smolensky, 1988] . There the same conceptual object, coffee, is represented by different hidden unit vectors, depending on whether the coffee is being thought about in the context of a cup, can, or spilled on the floor. Just as I have to alter my bodily configuration in order to continue to indicate the airplane as its position changes, so also the network must alter its hidden unit activations if it is to continue to represent coffee as the context changes.

This phenomenon is again a direct result of the fact that the locus of primary content ascriptions to PDP representations is at the whole content level. Since the representation/content relation is fixed only at the whole content level, the sub-total representation required for representing a particular partial content will, in general, vary from whole content context to whole content context, depending upon the other particular partial contents to be represented. This yields the many-to-one relationship.

The upshot of this second type of context-sensitivity is that typically, a network will not be able to adopt an adequate representation of a particular partial content in any arbitrary whole content context. The same constraints that require a change of representation will sometimes outstrip the representational capacities of the PDP system. Therefore, the content of such PDP representations cannot be represented in all contexts which (as we have seen) a unified, conceptual framework requires; the content in question must be non-conceptual. Although there is no reason why a symbolic representational system could not employ a many-to-one relationship (though it would be difficult to dream up ways to exploit the relationship), the fact that the choice of symbol to use is not dictated by the context means that this case of context out-stripping representational capability could never occur; conceptuality is guaranteed (as long as the context-invariant semantics for the atoms are justified, which they probably never are).[9]

In the context of the CNM, such examples are common. Just consider a network that

[9] Of course, there can be a many-to-one relationship between symbolic representations and *referents*; one *can* represent, in a production system say, the same block with either of the representations **block-57** or **The block X such that is-above(X, block-23)** (assuming that block 57 is the only block above block 23). But this would not be a many-to-one relationship between vehicles and *contents*; the contents of the two representations are different, even though their referents are the same.

is simultaneously learning two routes that happen to intersect at a few points. There is nothing that necessitates that the network use the same code for the place in both contexts; as a matter of fact, it might facilitate learning the sequences to have the representations distinct. Since the representations have to play different roles in the different route contexts, finding one representation that can do both might be difficult (but see the next chapter).[10]

7.4.5 Example 5: Sub-symbolic computation

Another reason for linking PDP with NCC falls out as a result of attempting to discover what it is that makes PDP representation an interesting representational genus, an alternative to symbolic representation. This is a question which Haugeland [Haugeland, 1991] has addressed. He argues that one might look in one of three possible places in order to find what it is that sets PDP representation apart:

1. the syntactic properties of the representations;
2. the relation between the representations and what they represent;
3. what the representations represent.

I would add a fourth place to look:

4. the contents of the representations.[11]

[10] Why isn't this case the same as the one in the previous note, i.e., a case of many-to-one vehicle/reference, but not many-to-one vehicle/content? Perhaps the best way to view these CNM representations is as having distinct contents, such as *place X in the context of route A* and *place X in the context of route B*? If one went all the way down this slippery slope, there would be no possibility of valid inference. For example, consider the inference from *the wall is grey* and *the wall is in front* to *there is something which is both grey and in front*. This is of the form $P(a), G(a); \exists x : P(x) \wedge G(x)$. By the reasoning being used against the many-to-one vehicle/content view of the CNM, one would have to conclude that the wall is being represented by distinct contents, since it is being represented in two different contexts. But this would mean the inference would be of the form $P(a), G(b); \exists x : P(x) \wedge G(x)$, which is invalid. We do not, in general, want to individuate contents as finely as the contexts in which they appear.

[11] Haugeland does talk about the "content" of a representation, but by this, he means what it represents, and not what I have been taking "content" to mean in this thesis.

I disagree with Haugeland that 3 is the place to look, since I feel that PDP and classical systems could very well be representing the same things, even if they are doing so with very different contents (this echoes my remarks in chapter 3 that we must keep the realm of reference fixed, and instead generalize the realm of content). And the connectionist community has, in general, avoided purely syntactic (type 1) distinguishing criteria (since symbolic representations can also be, e.g., continuous); rather, most settle on *distribution* as the important aspect of PDP representation, and the best recent analysis of distributed representation, by van Gelder [van Gelder, 1991], sees distribution as superposition, a type 2 criterion.[12] Perhaps contents are merely a relation between representations and the environment; perhaps 4 is subsumed by 2: these suppositions are interesting, but do not change the point. Which is: it does seem that the standard account, along the lines of 2, that is offered as a way of distinguishing PDP representation from other types, is inadequate. For example, one can construct symbolic cognitive architectures out of distributed, superpositional representations [Touretzky and Hinton, 1988, Smolensky, 1987]. The fact that one can use distributed representations to construct *non*-classical architectures only raises the question: what is it that unites the set of representations (that may or may not be a subset of the set of distributed representations) that are characteristic of viable, non-classical cognitive architectures? I think it is by answering *this* question that we can see what it must be about PDP-based architecture, *in addition to* distribution (for I do think that distribution of some sort is *necessary* for PDP to be interesting), that makes it a significant alternative to classical cognitive architecture.

Even if one admits that classical, symbolic representations can be distributed, it must be conceded that they would be of a very particular kind of distributed representation, a kind for which:

1. there is a conceptualized domain of objective values (objects, properties/relations, propositions, etc.) to be represented;

[12] What is superposition? Roughly, if R_1 represents i_1 and R_2 represents i_2, and R_1 and R_2 use the very same representational resources, then the representations of i_1 and i_2 are superposed; but the details are irrelevant here. The point is that superposition is a relation between representational vehicles and what they represent and is thus a criterion of type 2.

2. there is a scheme for assigning to certain configurations of the representational elements (symbols) these conceptual values as their contents;

3. the computational operations of the system only manipulate the elements in groupings that correspond to these symbols.

Understanding symbolic representation in this way, we leave open the possibility of sub-symbolic representation: representations for which, relative to an appropriate conceptual semantic interpretation (1), and despite the fact that there might be identifiable symbols (configurations of elements that correspond to a conceptual object or property) (2), there nonetheless exist computational operations for which there is no conceptual interpretation for them, nor for the partial content alterations that such operations make to the representations on which they act. For example, there might be no conceptual-level interpretation for an operation that changes the weights of the CNM predictive map after one epoch of learning. This is not to say that there is *no* level of description above the level of talk about numeric weight changes which can account for what's going on after each epoch of learning in the predictive map. Such activity is indeed intentional, with representational significance, and therefore has a characterization beyond that of mere mechanism. It's just that this characterization must be non-conceptual, for the reasons given.

Another way of seeing that the operations of the CNM are sub-symbolic, and therefore admit of non-conceptual interpretations, is to attempt to think of the D mapping as a predication: the assertion that something has the property of stimulating the agent in a particular way. Changes in the D mapping weights are therefore changes in what is predicated. But to what are these predications being applied? As we saw in the first example, they aren't objective, conceptual places. Thus, the change in predication cannot be understood as the retraction of some objective, conceptual predications and the assertion of others. The change has non-conceptual significance only.

The claim is that it is *this* property of PDP representations, that they are sub-symbolic, that they can underwrite a non-conceptual understanding of a system, which distinguishes them for the purposes of cognitive architecture construction. It is this sub-symbolism which sets PDP representations apart as an alternative to other forms of computation, at least for my (perhaps narrow) concerns.

Now we are in a position to see why distribution (at least in one sense of the term) is necessary for PDP if it is to provide an alternative cognitive architecture through sub-symbolic computation. Take distribution, not in van Gelder's sense of superposition, but in a content-based variation of an earlier notion [Hinton et al., 1986]. Call it "dual distribution": 1) several elements are involved in carrying each content and 2) each element being involved, at some point, in carrying several contents.[13] For purposes of illustration, consider the case of a system that has both conceptual and non-conceptual content. The possibility of sub-symbolic operations requires that the conceptual contents be represented by sets of elements (that is, it requires *many-to-one* distributed representation[14]); if all representations of conceptual contents (symbols) were atomic, then there could be no way for an operation to make a representational change that has no conceptual interpretation. Now it may seem that sub-symbolic representation does not, strictly speaking, require the converse: that each element be used in representing more than one content (i.e., that it does not require one-to-many distributed representation). But if this were the case, then sub-symbolic operations, in addition to not having any conceptual interpretation, might not have any meaningful interpretation at all. To see why, consider the following. If the system employs only many-to-one distributed representation and no one-to-many distributed representation at all, then any given element (or a given value for a variable) is involved only in the representation of one particular content. The representations for any two contents will be entirely disjoint. Thus, any representational operation that alters any proper subset of the elements of a content's representation will produce a non-content-bearing representation (which is really no representation at all). But such a system would not really provide an interesting alternative to symbolic representation, since, inter alia, any truly sub-symbolic operation would be rendered semantically void. Thus, one-to-many distribution is necessary after all, if the representational system is to have any interest. The

[13]This is different from superposition in that (at the least) it does not require all resources to be used in representing all items.

[14]This is not to be confused with many-to-one context-sensitivity, discussed earlier.

type of representation that PDP is offering as an alternative to symbolic representation is an alternative in that it comprises dual distribution *and* sub-symbolic operations.[15]

In closing, it should be pointed out that this exploration of the connection between PDP and NCC will still be of value even if, in addition to PDP's suitability, and the arguments of section 7.1 notwithstanding, some simple variations on classical architecture also turn out to be suitable for non-conceptual content. The value will remain, since in such a situation, one might still be able to appeal to other advantageous properties of PDP to isolate it as the cognitive architecture of preference. In any case, the work that has been pursued here, of finding a match between PDP and some kind of content is a *necessary* condition for it to be a successful cognitive architecture.

[15] van Gelder argues briefly (and, I think, successfully) that dual distribution is not sufficient for an interesting representational alternative. Something more, I claim, is needed. The notion of superposition that van Gelder puts forward certainly seems *compatible* with sub-symbolism, but I do not know if it is *necessary* for it, as dual distribution is.

CHAPTER 8
Cognitive maps & the development of conceptuality

8.1 A computational basis for the development of conceptuality

The foregoing showed how the sub-symbolic nature of connectionist computation could be employed to naturalize ascriptions of non-conceptual content. This chapter continues this thread by showing how some connectionist learning strategies can naturalize the transition from highly perspective-dependent contents to more systematic ones, and addressing some possible criticisms. In particular, it offers some specific means of determining the systematicity of the spatial representations in the CNM, so that transitions from perspective-dependence-reducing transitions may be identified. Furthermore, it identifies some of the conditions (tasks, training regimes, environments) under which such increases in conceptuality occur. After analysing the results of experiments that attempt to shed light on these questions, the chapter concludes by comparing and contrasting this work with related research.

8.2 Synthetic epistemology: Philosophy and AI/ALife

Sometimes in order to clarify the theories and concepts one would like to use to explain a natural system, it can be of great assistance to try them out on a simple, artificial system, which allows greater control and clearer analysis. Just as one might more readily come to a clear understanding of the principles of aerodynamics by studying a simple, artificial glider than by studying the particularities of the feathers and muscles of sparrows, so one might also see more readily the general structure of a proper psychology of real systems by first attempting to apply it to a simple, artificial agent.

Thus, to clarify some new ideas being proposed for the explanation of natural intentional systems, it seems a promising idea to turn to *synthetic epistemology*: the creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world.

Synthesis can thus be justified as an approach to understanding epistemology in the same way that it can be justified as an approach to understanding intelligence (AI), or biology (ALife):

> Artificial systems which exhibit lifelike behaviors are worthy of investigation on their own rights, whether or not we think that the processes they mimic have played a role in the development or mechanics of life as *we* know it to be. Such systems ... expand our understanding of life as it *could* be. By allowing us to view the life that has evolved here on earth in the larger context of *possible* life, we may begin to derive a truly general theoretical biology capable of making universal statements about life wherever it may be found and whatever it may be made of. [Langton, 1989, p.*xvi*, original emphasis].

The specific epistemological issue which this research addresses is understanding the nature of, and mechanisms underlying, the transition from heavily perspective-dependent to more systematic modes of representation. There are several reasons (given in chapter 7) for thinking that a connectionist architecture is much more suited than traditional symbolic architectures to the investigation of the development from less conceptual to more conceptual cognition. Furthermore, the acquisition of more and more sophisticated navigational abilities is plausibly seen as a paradigmatic case of the move from perspective-dependent to more perspective-independent ways of representing. Thus, the Connectionist Navigational Map (as explained in chapter 7) was developed for these purposes.

8.3 Simplicity as a virtue: arguments for and against the CNM approach

It might be thought that the grid "world" described in the last chapter is too impoverished to be of much interest. In particular, it might be thought that there are too few states, and that the "sensory properties" of each place are too coarse-grained, for this work to be of any relevance in understanding actual intentional systems. There are several reasons why I disagree.

First, this work can be seen as an extension of the research in learning finite state machines or formal grammars, e.g. [Cleeremans, 1992] and [Dienes, 1994]. The finite state machines in that work typically involve very few states with only one or two transitions possible from or to a state, and have no notion of the sensory properties of a state that may be shared with another state, and be sensed by the agent. I believe that by adding the complexity found in the CNM world, one begins to justify talk of learning spatial representations, instead of mere arbitrary grammars. But even if that assumption is illicit, the CNM paradigm should still be valuable, at least within the context of finite state machine learning.

Furthermore, coarse-grained sensations actually support the intended spatial interpretation of the CNM's activity. Since there are so few (i.e., 16) sensory signatures a location might have, the CNM cannot rely, in achieving its predictive aims, on merely recording the superficial sensory contingencies, but rather is forced to learn the more abstract spatial structure of its environment. To exaggerate this effect, I did not even let the CNM use the current sensations as an input to its predictive map, but rather forced it to use only its own representations.

It could still be objected that this, too, is unlike human cognition. It could be claimed that the way that humans and other animals achieve most of their navigation is by learning associations between actual detailed sensations, and not by developing some more abstract topological representation. That is, organisms predict what comes next by *looking* and seeing where they are.[1]

It would be a mistake to think that because I am interested in understanding how cognizers are able to make transitions from less conceptual to more conceptual ways of representing the world, that I somehow think that the majority of cognition involves representations that are at the extreme conceptual end of this scale. In fact, I believe that there are many kinds of cognitive interactions with the world that *require* relatively unsystematic, non-conceptual ways of representing, if they involve any representation at all. Furthermore, it may be impossible for any embodied, finite system to ever achieve total conceptuality or total systematicity. Nevertheless, I do think that there are interactions for which the ability to increase systematicity is a cognitive virtue, and spatial navigation is one of these.

[1] Thanks to David Rumelhart for pointing out this objection.

In order to pump your intuitions concerning these matters, consider the kinds of mistakes we (and other animals) do and do not make in navigating. Suppose that I leave the lecture theatre between talks at a conference in Sweden. While I am out of the room for a few minutes, the rest of the participants redecorate the lecture theatre so that it very closely resembles another one, with which I am familiar, in Brighton. They then hide (so as to give no clues about the true location of the theatre), and watch what I will do from behind doors, desks, etc. They might expect to have a good laugh upon seeing my puzzled expression, but they would *not* expect me to actually think I have somehow travelled hundreds of miles back to Brighton! Behind my puzzled expression is the thought "Why does this place look so much like the lecture theatre in Brighton all of a sudden?", not the thought "How did I cross the North Sea all of a sudden?!"

The objector might not find this story relevant, since what is being denied is that two different locations ever could, other than in the laboratory, ever yield *identical* sensations. On this view, the reason why I am not tricked into thinking I am in Brighton is due to my (perhaps sub-conscious) ability to make very fine sensory discriminations (e.g., I can see that the walls in the Swedish room have been recently painted to be that beige colour, whereas the paint is much older in the Brighton room).

But surely this is unlikely. It would imply that we would have difficulty recognizing the lecture theatre we are in now as the one that we were in before the break given that, e.g., the overhead projector has been moved slightly. We would be like the mnemonist S., whose eidetic memory made it difficult for him to recognize a face as the same as one seen earlier if the face's expression was different [Luria, 1968, reported in Glass & Holyoak, 1986, p. 330]. But we are not typically like that. Usually, one can recognize a place as being the same, in virtue of its relational properties, even though its intrinsic (sensory) properties have changed considerably since one's last visit. This cannot be explained by a model that does not allow for some topological, spatial representation, in addition to sensory association.

One might wonder why I am demanding that the CNM *learn* its spatial representations. Since the structure of space does not change within a creature's lifetime, surely the system of spatial representation could be innate. This may be, but there are two reasons that remain for using the CNM. First, in order to naturalize our systems of mental representation, not only do we have to have a (synchronic) understanding how they can be realized in our current physical structure; we must also have a (diachronic) understanding of how such abilities could be the product of a natural selection process [Cussins, 1992] (see chapter 2, section 2.4.5). So if there is no "development of spatial conceptuality" story to be told for any individual, then there must be some such story to be told for cognizing *species*. Second (but more concessively), there might be good design reasons for having an adaptive spatial representation system, even if its parameters are initialized at birth to some near-conceptual value.

Also, it should be re-emphasized that the CNM is for *synthetic* epistemology, and is obviously not meant to be a detailed model of the actual mechanisms of spatial learning in any natural system. Rather, it is meant to explore and illustrate some general principles and phenomena that are relevant under certain conditions (e.g., those in which local sensory information is not sufficient to guide navigation).

Nevertheless, I do plan to improve the CNM's "world" in several ways, including making the space continuous; using unit-free, routine-based actions; making the environment dynamic; and eventually using a real, non-simulated robot in a real-world environment. It is hoped that after some more general observations like the ones expressed in this chapter, these added degrees of realism will allow me to address more specific issues, in addition to further testing conclusions already drawn on the basis of these simpler "grid world" experiments.

8.4 Conceptuality as an effect of maximizing predictive success

Because of the feedback inherent in simple recurrent nets, the CNM's representations (location codes) change over time: at any time, a code may be used to represent a location or set of locations different from the location(s) that it is used to represent at a different time; and at any time, a given place may be represented by code(s) that are different from the ones used to represent that place at other times. The idea behind this research is that in some cases this dynamic process may be seen as a developmental one, in which the CNM's codes become more and more systematic, conceptual and perspective-independent.

In general, the CNM uses a different location code after each move, even when moving to a place that it has been before. That is the CNM typically uses different location codes on different occasions for the same objective location. Thus, typically, CNM representations are *non-systematic*. For our purposes, systematic[2] representation can be defined as follows:

> **Definition:** A system represents a location l systematically if there is a representation a such that:
>
> 1. whenever the system uses a, or a representation very functionally similar to a, it does so to represent l and not some other location l'; and
> 2. whenever the system needs to represent l, it is capable of using a, or a representation very functionally similar to a, to do so.

For the case at hand, these requirements boil down to:

> The CNM represents a location l systematically if there is a location code a such that, normally, a is active on the "current location" units if and only if the agent is currently at l.

Often, when speaking about the CNM's representations, I use expressions like "the same representation" or "different representations", when, strictly speaking, there is no such relevant issue of representational *identity*, but rather only representational *similarity*, in particular functional similarity. Thus, the above requirement is that normally all the codes a that the CNM has active on the "current location" units when at A are functionally very similar, and the CNM never has a code b, that is functionally very similar to one of the a, active on the "current location" units when the CNM is at a place other than A. Thus, there are at least three ways in which systematicity is a matter of degree:

1. the greater the number of different ways of getting to the place A that yield a code functionally equivalent to a, the greater the systematicity;

2. the greater the number of ways of getting to places other than A that yield a code functionally equivalent to a, the less the systematicity; and

[2] The terms "systematic" and "systematicity" have already been used, even (or especially) in the philosophy of connectionism literature; I am using it here as a technical term only. I think my notion of systematicity is related to the other notions of systematicity that have been used, but I am not yet clear on what exactly that connection is. However, this connection need not be made clear in order for the notion to do its work here as a measure of the degree of conceptuality which a set of representations exhibit.

3. the degree of systematicity will vary with the degree of functional equivalence in the above two conditions.

Therefore, the CNM, like a typical connectionist system, uses analog representations, rather than digital representations which can be *exactly* functionally equivalent. A consequence is that it seems unlikely that the CNM could ever achieve 100% systematicity, since it is a non-linear system, and even slight differences in location codes will, through iteration in the T mapping, most likely result in a large divergence at some point. Even if this is the case, it is not necessarily an argument against the CNM as a cognitive model, since it is not clear that, without external symbol systems, humans can be completely systematic either.

8.4.1 The development of conceptuality: another PDP/NCC affinity

Whereas in chapter 7 the connections between CNM representations and non-conceptual content were stressed, in this chapter the possibilities for extending the non-conceptual analysis of the CNM to the case of increasingly more conceptual intentionality will be examined. In example 4 from chapter 7, we saw that having many-to-one context-sensitivity might be necessary, given how difficult it is to find one representation that can represent the same thing in multiple contexts. But if a network can find such a common representation, then the conceptuality and systematicity of its contents will be increased dramatically. Such a situation is depicted in figure 8.1.

Selected locations on the route are labelled with each location's actual associated sensation vector (top four bits), the sensation vector that the agent (using the predictive map) predicted that the location would have (bottom four bits), and the code that the T mapping produced for the location (four numbers in between). Note that location (4 2) is represented by two diagrams (in the dashed box), one for each of the two ("before (5 2)" and "after (5 2)") contexts. Location (4 1) has two codes depicted: the top one is produced by travelling from (4 2) via (5 2) (solid lines); the (functionally equivalent) bottom one is the result of moving from (4 2) to (4 1) directly (dotted line).

At one point in the route learned by the agent in our example (solid lines), the trail doubles back. That is, the route involves going to (4 2) from the north, going east to

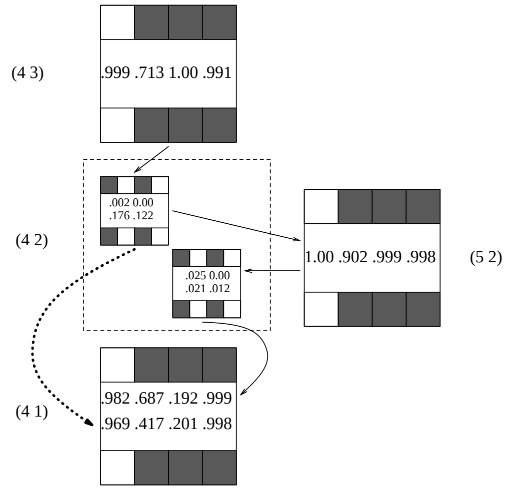

Figure 8.1: An illustration of how an increase in the conceptuality of location representations yields generalization in the CNM.

(5 2), going west back to (4 2), then going south to (4 1). The thing to note is that the first time at (4 2), the network uses the code **.002 0.00 .176 .122** to represent the location, while the second time there it uses the different, but very similar code **.025 0.00 .021 .012**. Although the network, at earlier stages of development, used functionally distinct vectors for this purpose, the network has, through learning, forced these two representations together to such an extent that they are functional equivalents in the local context.

Thus, it appears that the CNM has developed a *more* systematic representation for a place (i.e., it uses the same location code, across different contexts, for the same location). This systematicity yields a kind of generalization. Starting with the code for (4 3), the T mapping produces the first code for (4 2). But since this is functionally equivalent to the code used for the second appearance at (4 2), the network gives the correct result (in terms of both sensation *and* location vectors) when the agent moves *south* from the first (4 2) context (dotted line), instead of east, as it has always done before. A form of spatial generalization has occurred; the CNM is more than just a means of memorising a list of action/sensation sequences.

In section 7.4.1 of chapter 7, three constraints on a notion of objective space were enumerated. One of the three was:

3. For every simple movement m there is an inverse m^{-1} such that $T(T(a, m), m^{-1}) = a$

If it is claimed that figure 8.1 illustrates a case in which there has been an increase in the objectivity of spatial contents, one might expect it to illustrate a corresponding increase in degree to which these three constraints are met. It does indeed; meeting 3) is another consequence of the convergence of many-to-one representations into a functional cluster. Since the "before (5 2)" and "after (5 2)" representations have merged into the same functional cluster, we now have 3) holding for a =(4 2), m = **move-east**, and m^{-1} = **move-west**. Thus we now have a form of systematicity in which the representation one has after applying a move and its inverse is the representation with which one started.

In chapters 2 and 3 I argued that the development of intentionality may itself require an intentional analysis. I also cited evidence that suggests that this developmental process may have a complex, fine-grained structure, rather than being manifested in a single, inexplicable leap from mere mechanism to complete conceptuality. Thus, the appropriateness of an architecture for grounding this intentional characterization will depend upon its ability to march in step with the fine shades of increasingly perspective-independent contents that best characterize the development of conceptuality. I think the example just considered suggests that this is another reason why PDP architectures, such as the CNM, and non-conceptual content are a natural match.

Previously, I took such cases as demonstrating the emergence of systematic codes, which in turn suggested that the CNM was able to make transitions from less conceptual to more conceptual ways of interacting with the world. These, along with the counterpart transitions from more conceptual to less conceptual ways of representing, are the operations I take to be at the heart of cognition. However, I have since realized that the above demonstration needs to be qualified in two important ways.

8.4.2 Sameness of location vs. mere sameness of sensation

First, recent work [Holland, 1994] has pointed out that one must take care in inferring an increase in conceptuality on the basis of the kind of evidence just presented. Holland reports that the phenomenon of convergence of codes that correspond to the same place can be reproduced very reliably. But he makes an important observation: *the convergence* *also occurs for location codes that do not correspond to the same place, but merely to places that have, e.g., the same sensory properties.*

This is a consequence of the CNM's non-symbolic form of representation. In the last chapter it was pointed out that a key difference between symbolic and non-symbolic architectures is that in the former, associating a representation a with another representation d does not constrain the class of representations that the system can associate with a different representation, b. However, in non-symbolic architectures like the CNM, mapping a set of sensations to a via the D mapping *does* constrain what D can map to b. For example, if a and b are very similar (but still, say, functionally distinct), it is very difficult for D to map different sensation vectors to a and b. Looked at the other way, if it is a constraint on the codes that the CNM develops that D must map a and b to the same outputs, then the CNM will tend to develop similar codes for a and b.

Thus, it appears that the CNM's observed tendency to develop common codes for the same place encountered in different contexts can, at least in some cases, be explained as just a fortuitous by-product of a more pervasive, and less impressive tendency: to develop common codes for places that have the same descriptive mapping. If so, then this calls into question the appropriateness of the CNM for studying the development of more conceptual representations.

8.4.3 The functional equivalence of hidden representations

Also, the earlier study places too much emphasis on the actual Euclidean similarity/identity of two location codes. Conceptuality does not require that the location codes used to represent the same place in different contexts be themselves the same, or even similar; rather, they need only be *functionally equivalent*.[3] Conversely, the fact that two location codes are, e.g., clustered together in a cluster analysis, does not guarantee that the codes will play the same, or even similar, causal roles in the network. Two codes a and b are said to

[3] The notion of functional equivalence here focuses on the similarity of the *effects* of two codes. If, in addition, one paid attention to the similarity of the *causes* of the two codes, then one might not be able to distinguish identity and functional equivalence. I think that it is best *not* to include causal origins in a characterization of the functionality of a representation, so I am therefore compelled to acknowledge the difference between brute vector similarity and functional equivalence.

have a (second-order[4]) functional equivalence of $F_d(a,b) = -\frac{F_2(a,b)+F_1(a,b)}{2}$, where:

> $F_2(x,y) = \frac{\sum_{m=1}^{A} ||D(T(x,action_m)) - D(T(y,action_m))||}{A}$;
>
> $F_1(x,y) = ||D(x) - D(y)||$;
>
> A is the number (in our case 8) of actions available, and $action_m$ is the mth element of the list (N, NE, E, SE, S, SW, W, NW).

This value can be thought of as the negative average distance between corresponding sensory predictions (corresponding to the current and eight surrounding locations) that a and b give rise to.

There are two other ways of measuring functional equivalence that have been considered: the percentage F_p of neighbouring output *pattern* predictions that are the same, and the percentage F_b of neighbouring output *bits* that are the same. That is:

> $F_p(x,y) = 100\frac{\sum_{m=1}^{A} P[D(T[x,action_m]), D(T[y,action_m])]}{A}$; and
>
> $F_b(x,y) = 100\frac{\sum_{m=1}^{A} B[D(T[x,action_m]), D(T[y,action_m)]}{A}$;

where:

> $P(d_{xm}, d_{ym})$ is 1 if the (thresholded) sensory predictions d_{xm} and d_{ym} are equal, and 0 otherwise; and
>
> $B(d_{xm}, d_{ym})$ is the Hamming distance between sensory predictions d_{xm} and d_{ym}.

The former measure is more strict than the latter; two codes a and b may be such that $F_b(a,b) = 75\%$, yet $F_p(a,b) = 0\%$ (i.e., they may always disagree on, say, bit 1 of the 4 possible output pattern bits, but agree on all others). Its advantage is that it better avoids apparent functional equivalences that are actually spurious in that they depend on some accidental similarities (e.g., those that are a product of the contingent distribution

[4] Obviously, this definition of functional equivalence can be generalized via a recursive definition to nth order functional equivalence (i.e., the negative of the distance between predictions made for locations up to n moves away) for arbitrary n. For our purposes it is sufficient to use only these first few terms for such a generalization, since they dominate the results.

of sensory properties in the environment) between the output that a is producing and the output that b is producing. But the latter may be a better measure for some purposes, since agreeing *somewhat*, if not perfectly, on neighboring predictions, indicates some degree of functional equivalence that may be of some explanatory use (especially if the non-sensory properties of a place – presence of food, danger, etc. – are reliably correlated with its sensory properties). In the 75%/0% case, above, it seems that b and a have *some* degree of functional equivalence, even if it is systematically distorted. Thus, all three measures were used in reporting the results of the experiments below.

It may sometimes be useful to acknowledge the fact that two location codes are functionally similar with respect to the actions that the network actually made at those locations, while being functionally divergent with respect to the the remaining actions, which are, in effect, "don't care" values as far as the training regime is concerned. This notion of *relevant* functional divergence is denoted by F^*, and is calculated by only using actions that have actually been taken by the agent in the involved location(s) when summing and normalizing in the first equation above. It is desirable, of course, that $F(x, y)$ be low for any co-referring x and y, but such a situation is not strictly required for the corresponding $F^*(x, y)$ to be low, which would in itself constitute a degree of systematicity. In the experiments reported here, I ignored this weaker notion of functional equivalence, since one of our main interests is in the generalization from what has been explicitly experienced to what has not.

Note that functional equivalence of any of these three kinds is independent of *correctness*: two codes may give rise to the same predictions (and thus have $F_d = 0$ and $F_p = 100\% = F_b$), yet both may be completely wrong in those predictions. The connection with correctness will be captured in two ways in the experiments that follow: the criterion that the net learn until it correctly predicts all sensations on its route; and the generalization that will naturally result in the cases of high systematicity.

In the experiments that follow, I give an example of the cases that justify the introduction of these functional equivalence measures: cases in which Euclidean distance/clustering would suggest a functional equivalence that is not present, and cases in which Euclidean distance/clustering would suggest functional divergence that is not present. This aspect of the research, then, can have a relatively broad application, even if one is not interested in synthetic epistemology, connectionist navigation or the development of conceptuality.

8.5 Requirements for systematicity: an hypothesis

In order to address these issues concerning the requirements for the development of conceptuality, a hypothesis was formed concerning the conditions under which this style of representation will arise in the CNM, and experiments have been conducted to test this hypothesis.

Given the definition of systematic representation in section 8.4, the central hypothesis of this chapter can be stated thus:

> **Hypothesis:** The CNM will only develop a systematic representation of a location l if its encounters with l, and with locations that resemble l, are so structured as to make such a form of representation a useful means of minimizing the error of its predictions.

The plausibility of the hypothesis is a consequence of the CNM's non-symbolic form of representation, as discussed in section 8.4.2. The holistic, as opposed to atomistic, nature of representation in the CNM implies that systematic representation will not be the default. Since what *primarily* determines whether the two location codes used at two different points in a route are similar is the similarity of the sensory predictions that such codes are required to produce (and not the identity of the two locations in question), the CNM will tend to violate the first of the two requirements for systematicity. Thus, it is only *likely* to satisfy the first requirement if its routes through its environment which generate its training regime are structured in particular ways.

The hypothesis itself doesn't have much force without some specifics concerning what kinds of structure the CNM's encounters must have in order to make the hypothesis true. If one prefers, one can rephrase the hypothesis into a question: what kind of spatial behaviours, if any, compel the CNM to form systematic spatial representations?

I attempted to answer this question by considering it for each of the two components of the working definition of systematic representation:

1. under what conditions does the CNM avoid allocating functionally equivalent codes

to distinct locations (even though the locations, e.g., have the same description)?; and

2. under what conditions does the CNM succeed in using functionally equivalent codes for the same location in different contexts?

In trying to answer these questions, one major obstacle to systematic representation, already alluded to, must be understood. If the CNM needs, in two different contexts A and B, to produce the same (or very similar) outputs on the D mapping, then there will be a tendency for it evolve weights such that the codes that are active in those two contexts, a and b, are functionally equivalent, even if the CNM is at different (albeit sensorily similar) locations in those two contexts. Thus there is a tendency to violate the first of the two requirements for systematic representation. In what situations, if any, can this tendency be overcome, such that systematic representations *are* developed?

But this is only one example of how the predictive demands placed on the CNM constrain the kinds of representations used. Another example is that making the same move at two different parts of the route will tend to produce similar codes for the location after those moves. The representational demands of a recurrent network are extremely holistic, with the "optimal" representation for the current situation being determined both by what it will give rise to in the arbitrarily distant future, and by what what gave rise to it in the arbitrarily distant past, in addition to the constraints of the present. Not only does the code that is used for the current location have to be mapped to the current sensations via the D mapping, but it needs to give rise to a code that can lead to the right predictions for the next step in the route, and it needs to be such that it can be the product of inputting the last code and action into the T mapping.

8.6 Principles & Predictions

To make substantive the hypothesis of the previous section, I used it to make some predictions concerning the conditions under which systematicity would and would not develop.

First, I noted four principles that I take to characterize the holistic interdependence of CNM representations (i.e., the aspects of the CNM that make it non-symbolic, as discussed in the previous section and section 8.4.2):

1. same inputs tend to produce same outputs
2. different inputs tend to produce different outputs
3. same outputs tend to require same inputs
4. different outputs tend to require different inputs

These are, of course, only rough guides and tendencies, which are defeasible. But in the context of the CNM, I appealed to the above principles to suggest some more concrete tendencies concerning the functional equivalence of location codes that the CNM develops.

I focussed on the case of codes that the CNM develops to represent *sensorily equivalent* places. This is because we are here interested in two kinds of case: the divergence between codes that represent sensorily equivalent but spatially distinct places, and the equivalence between codes that represent the same place (which, obviously, must also be sensorily equivalent).

I used the principles to derive the following postulates[5], expectations concerning how the CNM's codes would develop (numbers in brackets indicate which of the principles were used to derive each postulate):

1. $D(a) = D(b) \rightarrow a = b$ [3];
2. $ma_{-1} = mb_{-1} \rightarrow a = b[1]; ma_{-1} \neq mb_{-1} \rightarrow a \neq b$ [2];
3. $D(a_{-1}) = D(b_{-1}) \rightarrow a = b[3]; D(a_{-1}) \neq D(b_{-1}) \rightarrow a \neq b$ [4];
4. $ma = mb \rightarrow a \neq b[4]; ma \neq mb \rightarrow a = b$ [3]
5. $D(a_{+1}) = D(b_{+1}) \rightarrow a = b[3]; D(a_{+1}) \neq D(b_{+1}) \rightarrow a \neq b$ [4];

where:

[5]At least one or two of these postulates seem to have an analogue in [Cleeremans, 1992, pp 64-65] (e.g., postulate 5).

"=" means "similar" for movement and sensation vectors, but means "functionally equivalent" for location codes; and

"$\rightarrow$" means "tends to make true".

Postulate 4 requires some explanation, since it does not hold unconditionally. In general, the similarity or difference of moves made from a and b has no implication in itself for the functional equivalence of the codes. But it does have implications when interacting with other contexts. In particular, if $D(a_{+1}) = D(b_{+1})$, then $ma \neq mb \rightarrow a \neq b$. This is because differences in a and b will be required in order to cancel out the differences in ma and mb in order to have a constant result.

Conversely, if $D(a_{+1}) \neq D(b_{+1})$, then $ma = mb \rightarrow a \neq b$, by principle 4. To see why, first note that principle 4 implies that $D(a_{+1}) \neq D(b_{+1}) \rightarrow a_{+1} \neq b_{+1}$. Next, note that there will be an even stronger push (via principle 4 again) for $a \neq b$ than there would be based on prediction 5 alone, since the similarity in the moves ma and mb must be compensated for by greater differences in a and b in order to achieve a comparable difference in a_{+1} and b_{+1}. There will be no special tendency produced by $ma \neq mb$.

In stating these tendencies, my use of "=" and "$\neq$" suggests that I am once again assuming either completely equivalent or maximally different description vectors. But in fact, the relevant description and movement vectors may be more or less similar or different. These differences should affect the functional equivalence of the relevant location codes accordingly, but given a random distribution on sensation vectors and moves, I believe these additional modifying factors can be ignored in the analysis.

In light of these postulates, I defined nine (non-exhaustive) types of route, or scenarios, that I thought might generate a large variation in the degree of systematicity of the representations the the CNM develops for two locations that are sensorily equivalent. The situations are listed in figure 8.2.

Using the five principles, I predicted the following rough ordering of these situations with respect to the degree of systematicity that they impose on the CNM's representations for the two locations, from most systematic to least:

SIDO These scenarios should yield the best systematicity, since because functional diver-

DIDO : Different ways in, different ways out. The route that the CNM takes approaches each place from several different directions, and leaves from each place in several different directions.

SISO : Same way in, same way out. There are four possible sub-cases:

- SS both the single direction in and the single direction out are the same for the two locations
- SD the single way in is the same, but the single directions out are different for the two locations
- DS the single ways in are different, but the single direction out is the same for the two locations
- DD both the single ways in and the single directions out are different for the two locations

DISO : Different ways in, same way out. For each location, the CNM's route approaches from several different directions, but always leaves by the same direction. There are two sub-cases:

- S the single way out is the same for both locations
- D the single way out is different.

SIDO : Same way in, different ways out. For each location, the CNM's route approaches from one direction only, but leaves by several different directions. There are two sub-cases:

- S one in which the single way in is the same for both locations, and
- D one in which the single way in is different.

Figure 8.2: The classification of routes used in the experiments.

gence between the codes for different places is fostered by exploring the different sensory surround of the two locations, yet each of the two locations is entered via a constant approach, providing a basis for the development of very similar codes for the same place. Within this group SIDOD should be more systematic than SIDOS, since the differing ways in to the two locations will add the the divergence between their location codes.

DIDO This should be next best with respect to systematicity, because although the lack of a common approach to the locations will yield a divergence between the codes used for the same place, there will be a greater divergence between the codes used for the two different places, due to the exploration of their different sensory surround.

SISO These should yield poor systematicity, due to the lack of exploration of the two locations' different sensory surrounds. However, SISODS and SISODD should be more systematic than SISOSS and SISOSD, since the single moves in are not the same between the two locations, thus causing *some* functional divergence between the codes for the two places. SISODS should be slightly more systematic than SISODD, and SISOSS more than SISOSD, for reasons similar to the ordering given within the SIDO category, above. All of these should be more systematic than the DISO scenarios, since at least the codes for a location are being produced by a common factor: the move in.

DISO DISOS should yield poor systematicity, but DISOD should be even worse, since in DISOS there is at least one basis for forcing a divergence between the codes that represent the two places: the different predictions required of their common move out of those places (the same code cannot produce both 1111 and 1001 when combined with the move North). In DISOD, the single ways out are different for the two locations, and thus there will be no need to develop different codes to accommodate the different predictions (the same code *can* produce 1111 when combined with North and 1001 when combined with E).

Of course, the variables used in calculating these predictions are not all of the ones that are relevant in determining the degree of systematicity of the two representation codes. In particular, I have said nothing about the distribution of description vectors for the places surrounding the two locations in question, yet this will typically have considerable effect. For example, if there were local duplications (if, e.g., the locations surrounding a and b had corresponding description vectors), then postulate 5 would suggest that exploring the sensory surround of a and b will push the two codes together, not cause them to diverge, as was assumed in the rationales for the above predictions. Nevertheless, if one assumes a uniform distribution of sensation vectors, the predictions that I have made will tend to hold, given that local duplications are highly unlikely.

8.7 Experiments & results

To test these predictions, I had the CNM learn 7 routes, each route realizing a different route type (see figure 8.3). The particular environment that was used is shown in figure 8.4 (the four-bit binary vector at each location indicates the description or sensation vector associated with that location). The CNM converged on a solution with no errors within, on average, 16330 epochs of training.[6] The learning rate was 0.01, and the momentum was 0.5.

As I surmised (in section 8.4.3), the standard Euclidean measure of distance (and attempts at functional analysis based on it, such as cluster analysis) is an unreliable measure of functional equivalence. The non-linear nature of networks means that sometimes codes that are geometrically close will have different functional properties, and sometimes codes that are relatively geometrically distant will be functionally equivalent. An example of this was found in the codes (C_{29}, C_{24} and C_{33}) the CNM learned for the DIDO route (for moves 29, 24, and 33; see figure 8.5). Although the distance between C_{29} and C_{33} was less than than the distance between C_{29} and C_{24}, the functional equivalence of the former pair was less than that of the latter pair, on all three of the measures of functional equivalence.

[6] In a few of the simulations, there were a few prediction errors (at most 2 on any route) with respect to the learned route, but none of the errors involved the two locations under scrutiny nor their immediate neighbours.

1. **DIDO:** N; W; S; SE; E; N; N; W; S; E; S; S; NW; N; NW; E; N; S; E; SE; S; W; W; SW; NE; E; E; NW; W.
2. **SISOSS:** N; W; S; SE; SE; N; W; N; N; W; NE; SE; SE; S; SW; N; W; NE; W; N; W; SE; S; SE; N; W; W; NE; N; W; S; S; SE; E; N; W; N.
3. **SISODD:** N; E; S; S; W; NW; E; N; E; E; S; W; S; W; S; NW; NE; N; E; SW; E; S; W; N; N; E; S; S; W; N.
4. **DISOS:** N; W; SE; E; S; W; NW; N; E; W; SW; SE; E; E; W; NE; N; W; W; S; SE; S; E; N; W; N; NW; NE; S; W; SE; E; SE; W; W; N.
5. **DISOD:** N; E; S; S; N; N; W; E; SW; S; E; N; N; NW; SW; E; E; SE; S; W; N; W; NW; NE; S; E; S; SW; SE; N; N; W.
6. **SIDOS:** S; E; N; W; NW; E; N; SE; SW; S; E; E; NW; W; NW; E; E; SW; S; E; S; NW; N; NW; E; S; S; E; W; N; NW; E; W; SE.
7. **SIDOD:** S; E; E; NW; W; N; W; SE; S; E; S; NW; N; N; E; S; SW; E; W; N; N; S; S; E; N; W; N; N; SW; SE.

Figure 8.3: The route types used in the experiments, and the particular move sequences that realized them

	2	3	4	5	6
2	0110	0010	1111	0111	0111
3	1000	1000	1001 A	1010	0110
4	0001	1000	1100 HOME	0100	0101
5	1011	1010	0010	1001 B	1000
6	1110	1111	0010	1001	1111

NORTH

Figure 8.4: The region of the grid world used in the experiments.

Code 1	Code 2	Distance	F_b	F_p	F_d
C_{29}	C_{24}	0.87	87.50%	62.50%	-1.056
C_{29}	C_{33}	0.74	84.38%	50.00%	-1.517

Figure 8.5: An example of Euclidean similarity and functional equivalence coming apart.

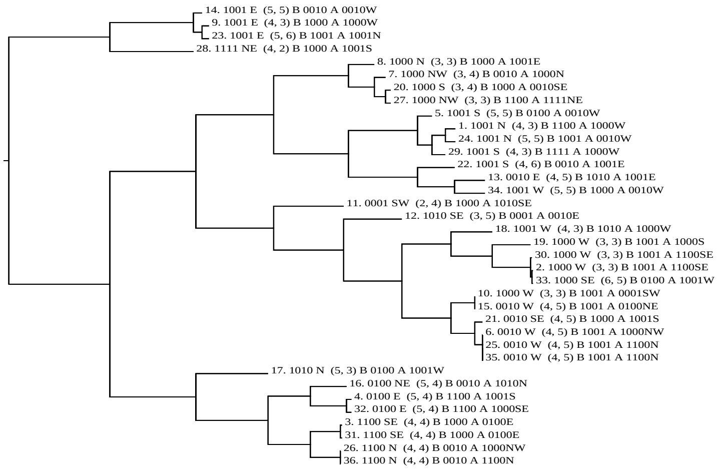

Figure 8.6: Cluster analysis of all location codes used in the DISOS route.

8.7.1 Qualitative analysis

One can use cluster analysis to get a rough idea of the different degrees of systematicity developed in learning the different types of routes. Figure 8.6 shows the cluster analysis of the location codes developed in learning the DISOS route. Labels indicate, respectively: the move number that produced the code, the description vector for the location, the move made, the coordinates of the location, the description vector of the previous place, the description vector of the following place, and the move taken to get there. Note how the codes corresponding to (5, 5) are found in several parts of the tree, suggesting low functional equivalence between them. The same applies to the codes for (4,3). Note also that codes for (5,5) and (4,3) are often clustered together, suggesting a high functional equivalence between them. Both of these factors indicate a very low degree of systematicity.

In contrast, the cluster analysis of the codes developed for the SIDOD route (figure 8.7) suggests a high degree of systematicity. The codes for (5,5) are all clustered together, as are the codes for (4,3), and the (4,3) and (5,5) codes are in different (albeit neighbouring) sub-clusters, suggesting that they might be functionally divergent, despite the sensory equivalence of the two locations.

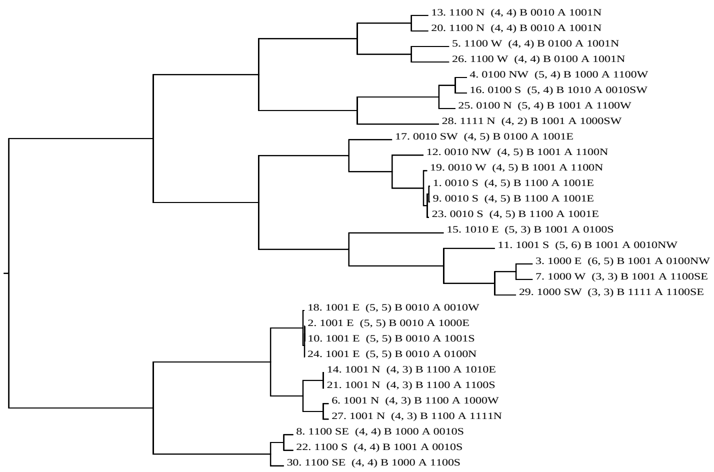

Figure 8.7: Cluster analysis of all location codes used in the SIDOD route.

8.7.2 Quantitative analysis

However, in order to provide a more detailed comparison of the systematicity of the location codes developed for each route type, we need a quantifiable measure of systematicity. In keeping with the two components of the definition of systematicity, systematicity should be maximized when the functional equivalence between codes for the same location is maximized, and when the functional equivalence between codes that correspond to different locations is minimized. Thus, systematicity can be seen as the average functional equivalence of codes for the same location minus the average functional equivalence of codes that represent different locations. For the particular cases considered in the experiments, this can be formalized as:

$$S(A,B) = 2\frac{\sum_{i=1}^{N}\sum_{j=i}^{N} F(a_i,a_j)+F(b_i,b_j)}{N(N-1)} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} F(a_i,b_j)}{N^2}$$

where:

A and B are the distinct yet sensorily equivalent locations the codes for which are under consideration (in our case, the A and B were the locations (5,5) and (4,3) in all routes);

N is the number of times that the route enters the places A and B (in our case 4);

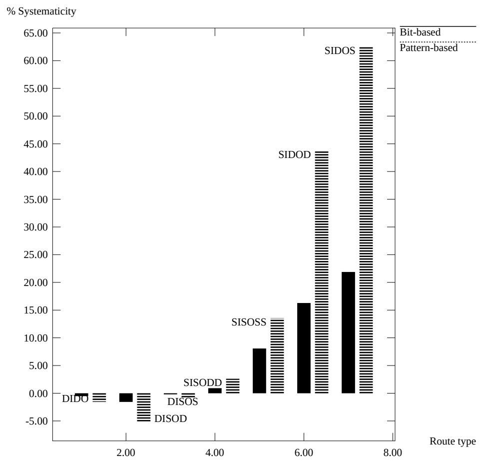

Figure 8.8: The observed bit- and pattern-based systematicity of the 7 tested route types.

F is the functional equivalence measure being used, be it F_p, F_b, F_d (see section 8.4.3); and

a_i and b_i are the location codes that are active on the ith visits to A and B, respectively.

The first term sums up the functional equivalences of the four codes that represent A and the functional equivalences of the four codes that represent B, and then divides by the number of such comparisons (in our case 6) to get an average; the second term sums up the functional equivalences of the codes that represent different places, and then divides this sum by the number of such comparisons (in our case 16) to yield another average. The difference then expresses the degree of systematicity.

The systematicity results using pattern- and bit-based functional equivalence measures are shown in figure 8.8. I also calculated the systematicity of the 7 scenarios using the distance-based measure, shown in figure 8.9 (note that this is not a measure of the Euclidean distance between the codes, but a measure of the distance between the sensations that two codes predict).

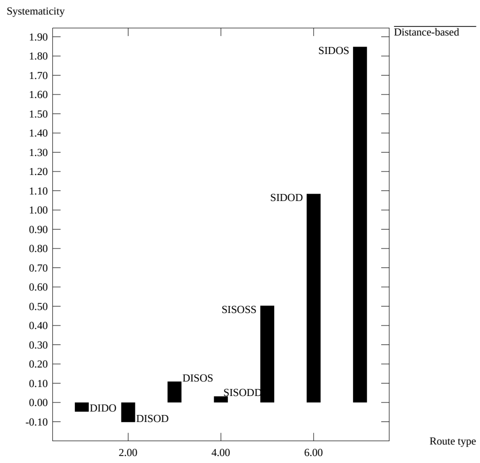

Figure 8.9: The observed distance-based systematicity of the 7 tested route types.

8.8 Discussion

The data are fairly univocal. The SIDO scenarios produce the most systematic codes, the SISO scenarios to a lesser extent, and the DISO scenarios even less. This agrees with my predictions, and thus supports the postulates, principles and hypothesis of sections 8.5 and 8.6.

However, two aspects of the predictions were not borne out. First, although the predicted general ranking of the SIDO, SISO, and DISO scenarios was correct, the predictions of the rankings of the sub-types within those groups were not. In particular, I expected SIDOD to be more systematic than SIDOS, and SISODD to be more systematic than SISOSS, yet the converse was found in both cases. It seems that having a common move in helps the codes for the two locations to diverge from each other, which suggests that in some cases principle 1 and postulate 2 (see section 8.6) do not hold. Further work, then, should include an investigation into that principle and postulate.

Second, I expected the systematicity of the DIDO route to be between that of the SIDO and SISO routes, but it is in fact near the bottom of the scale, better only than the DISO routes. This suggests that the commonality of the codes for the same location has a greater weight in the determination of systematicity than divergence between codes of different locations. This would also explain the negative systematicity results for the

DISO routes.

Note that, predictably, the high systematicity of the SIDO routes yielded perfect generalization. That is, since the CNM was trained until it got all the predictions on its routes correct, the four codes that it used for each of the four times in (5,5) each make a correct prediction for what would result from moving north, east, south and west, respectively. But since the CNM in this case developed a systematic representation of (5,5), these four codes are highly functionally equivalent, and thus a correct prediction is made if the CNM considers moving south from (5,5) at a point in the route where it normally would have gone east. This shows that the CNM is doing more than just memorizing a route (see section 8.4).

Another look at the cluster analysis of the SIDOD route (figure 8.7) suggests that another kind of generalization might be at work. Even though the SIDOD structure of the route was only expected to foster a systematic representation of (4,3) and (5,5), it appears that *every* developed location code meets the systematicity requirements (contrast the ordered grouping of co-referential location codes in figure 8.7 with the relatively jumbled groupings in figure 8.6). Perhaps developing systematic representations for a few locations can serve as a catalyst that bootstraps systematic representation in general, for locations that have not been the focus of a SIDO strategy. This will be the subject of future work.

8.9 Comparisons with other work

Independently of the work done here, there has been work on applying simple recurrent nets (SRN's) to the task of learning finite state automata (FSA's) which suggests their capability to develop systematic representations of those automata. In particular, the cluster analysis in [Cleeremans, 1992, chapter 2] of an SRN trained to predict the next letter in a sequence constructed from a simple grammar suggested that the net had indeed developed hidden unit patterns that were active if and only if the portion of the string processed so far corresponded to a particular node in the FSA representation of that grammar. Furthermore, Cleeremans went on to examine one parameter that seems crucial in determining whether or not such systematic representations will develop: the number of hidden units. Given too many hidden units, the network will use different regions of the hidden unit space to represent the same state, thus preventing generalization.

The work here can be thought of as complementing Cleeremans', in that it highlights external, rather than internal, constraints on the development of systematic representation. However, there are several other differences between that work and this,

Cleeremans used a very different training strategy. In order to train his network, he used 60,000 sequences with an average of 7 patterns per sequence, yielding 420,000 training patterns (as opposed to my use of 30 patterns or so). He thus had little idea of which of those patterns were crucial for systematic representation. The lighter approach here allowed me to investigate the relatively minimal requirements for the development of systematicity.

Also, the nature of the task is different: the CNM learns by predicting the sensory properties of states, while Cleeremans's model predicts possible sequences of letters, which are analogous to the *actions* in the CNM (i.e., they are what effect state transitions). Perhaps the CNM could be modified to use both kinds of learning to further constrain the development of systematicity. This would also improve the CNM's ability to aid in navigation, since it would be able to rule out some possible action sequences as being "ungrammatical". Correlations between these restrictions on movement and sensory properties could then be learned, yielding a deeper understanding of the causal properties of its environment.

In his analysis of (what I would call) the systematicity of his network's representations, Cleeremans relied only on Euclidean distance and cluster diagrams, which as has been shown do not always indicate the true functional equivalence of representations. However, since his behavioral tests were so exhaustive and high-scoring (e.g., his network correctly categorized 130,000 randomly generated strings as grammatical or ungrammatical), perhaps the high number of training patterns ensured a tight connection between distance and functionality.

Given that the development of spatial representations in the CNM can be thought of as the development of "location permanence" (see chapter 7, section 7.4.1), recent research into connectionist models of the development of object permanence [Mareschal and Plunkett, 1994] are highly relevant to the work done here, especially since the task in that work is (visual) prediction using an SRN. However, despite these similarities, there are some serious differences. For one, Mareschal & Plunkett's work has the advantage of successfully addressing and explaining actual human developmental data. Also, their evaluation of their network is entirely behavioural; i.e., they do not analyse the representations their network develops via cluster analysis or calculate systematicity measures.

Finally, there has been some investigation into the conditions under which SRN's learning FSA's can transfer what they have learned to other domains [Dienes, 1994], which can be seen as another kind of generalization. For example, it was found that in order to achieve transfer, a network should be presented with sequences that have repeated elements. Perhaps the principles that were useful here in predicting and explaining the conditions for the development of one form of generalization could be of use in explaining why repeated elements are so crucial to developing this other kind of generalization.

8.10 Future work

In addition to the future work already mentioned (cf sections 8.3, 8.8 and 8.9), some other possibilities should be mentioned.

The generalization exhibited by the CNM so far only involves different combinations of transitions that it has made before. Another important kind of generalization is to transitions not made before, but for which one has been given enough information to make a successful prediction. For example, suppose an agent using the CNM has never moved south from (4,4) to (4,5), but it has been to (4,5) via moving east from (4,4) to (5,4), then south to (5,5), and then west. It would be very significant if the CNM could develop representations so systematic that the code for "south from (4,4)" was functionally equivalent to the code for "east, south and west from (4,4)", even though it had never moved south from (4,4) before. Specifically, it might be useful to consider under what conditions the CNM develops codes such that the action vectors cause systematic movements in the principal component space of the location codes.

Perhaps the CNM is on its way toward this, as evidenced in the unexpected systematicity in figure 8.7. But another idea for how the CNM might achieve this is for it to have a small core of systematic location codes which it repeatedly redeploys in order to represent different areas. This might also address the scaling-up problems of SRN's that have been observed [Cleeremans, 1992, p 66]. It is unclear, however, how such a structure might be learned, so it is unclear to what extent such an architecture would be furthering a connectionist naturalization of epistemology, as opposed to assuming a complex innate symbolic mechanism.

Finally, if the CNM were to be refined so that it might be used to model and explain actual data of some sort, a natural area of application is the spatial learning of rats. The "place cells" [O'Keefe and Nadel, 1978] observed to be in the rat hippocampus, cells which are maximally active if and only if the rat is at a particular location, sound very much like the systematic location codes developed in the CNM.

CHAPTER 9

Conclusions and Conclusion

9.1 In brief

The first half of the thesis focused on content. I argued that that we already have to hand empirical data that demand a notion of non-conceptual content, I showed that this notion is a coherent one, and defended it against attack. The second half of the thesis focused on computation. There I defended computation's role in psychological explanation, and then argued that computation is needed for the specification of non-conceptual contents. Next I argued that the naturalization of non-conceptual content is facilitated by sub-symbolic computational architectures, such as the Connectionist Navigational Map.

9.2 In more detail

9.2.1 Content

Chapter 2 introduced the notion of a content-based explanation, and presented the Example, from the developmental psychology literature on object permanence in infancy. I argued that the Example is just one instance of the Recalcitrant Phenomena, which require content-based explanation, but which cannot be explained in terms of objectual content.

The route into an understanding of non-conceptual content passes through an understanding of conceptual content. Also, conceptual content accounts are the principal rivals to non-conceptual accounts. Therefore, I began chapter 2 with a careful discussion of conceptual content, including its connections to the Generality Constraint and objectivity. Armed with an understanding of what concepts are, non-conceptual content was stipulated to be content that has constituents which are not concepts; concepts are those constituents of content that are individuated with respect to their role in judgement alone. I showed that widely acknowledged facts concerning conceptual content can be explained by this stipulation. Non-conceptual content then turned out to be this: content composed of constituents which are individuated with respect to more than just judgement. Returning to the Example, I showed how this notion of non-conceptual content avoided the pitfalls that beset attempts to explain the infant's behaviour in conceptual terms. I contrasted my notion of non-conceptual content with those of other writers; on my view, non-conceptual content can be the object of attitudes such as belief, and it can have determinate truth-conditions.

In chapter 4, McDowell's recent criticism of non-conceptual content was discussed. I gave several reasons for rejecting his conclusion that all content is conceptual. The most portable insight was this. It is true that in understanding others, we cannot take a sideways-on view of their relation to the world; we must understand them by sharing their perspective. But it is sufficient, and necessary, for sharing their perspective that we share a unitary referential realm. It is not necessary for us to share a unitary realm of content. That is, although we may be required to individuate and ascribe non-conceptual contents "from without", this does not mean we are taking a sideways-on view in any illicit sense.

9.2.2 Computation

Content-based explanation requires a way of practically yet canonically specifying the contents to which it adverts. Chapter 5 argued that conventional "that"-clause specifications of conceptual content will not work for non-conceptual content, and investigated various alternative means of specification. The most promising candidates shared the properties of being embedded and embodied: they made essential use of the abilities which realize the content, and the context in which the content is normally entertained. It was argued that computational notions are required for providing such specifications.

Thus, with one reason why non-conceptual content explanations are facilitated by the use of computation established, chapter 6 defended the coherency of computational explanations in cognitive science against recent attacks from Putnam and Searle. Their objections centre on the claim that any physical system could be understood to be realizing any formal computation. My response was to admit that although a richer notion of computation is required in cognitive science, even the formal notion may play a part, and is not vacuous. Even formal computation includes a causal requirement which dispels the universal-realisation worries of Putnam and Searle in a non-relativistic way.

With the ontological status of computation back on a safe footing, chapter 7 looked at another reason why non-conceptual content explanations are facilitated by the use of computation: naturalization. I argued that the best computational vehicle to naturalize non-conceptual explanations is a sub-symbolic one. I did this in two ways. First I pointed out that arguments such as Fodor's, which link conceptual content to classical architecture, thereby establish a link between non-conceptual content and sub-symbolic architectures. Second, I described a particular case of a connectionist sub-symbolic system, the Connectionist Navigational Map (CNM), which I designed as a way of investigating and illustrating the affinities between connectionist representations and non-conceptual content.

Chapter 8 continued the use of the CNM as an expository device. The first part of the chapter showed how some connectionist learning strategies can naturalize the transition from highly perspective-dependent contents to more conceptual ones. Then some possible criticisms were addressed, by offering some specific means of determining the systematicity of the spatial representations in connectionist systems, so that perspective-dependence-reducing transitions may be identified compared. Furthermore, it identified some of the conditions (tasks, training regimes, environments) under which such increases in conceptuality occur.

The philosophical considerations of the thesis show that the concept of non-conceptual content is coherent; the computational and psychological considerations show that it is explanatory.

Bibliography

Allen, C. (1992). Mental content. *Brit. J. Phil. Sci.*, 43.

Anderson, J. (1983). *The Architecture of Cognition.* Harvard University Press, Cambridge, Mass.

Baillargeon, R., Spelke, E., and Wasserman, S. (1985). Object permanence in 5-month-old infants. *Cognition*, 20:191–208.

Bates, E. and Elman, J. (1992). Connectionism and the study of change. Technical Report 9202, Center for Research in Language, University of California, San Diego.

Blackburn, S. (1994). Content. In *The Oxford Dictionary of Philosophy*, page 79. Oxford University Press, Oxford.

Brewer, B. (1987). *Strawson and Evans on Objective Particulars and Space.* University of Oxford. B.Phil. thesis.

Burge, T. (1979). Individualism and the mental. *Studies in Metaphysics*, 4. Midwest Studies in Philosophy.

Burge, T. (1982). Other bodies. In Woodfield, A., editor, *Thought and Object: Essays on Intentionality*, pages 97–120. Clarendon Press, Oxford.

Campbell, J. (1985). Possession of concepts. *Proceedings of the Aristotelian Society*, 85:135–156.

Campbell, J. (1994). *Past, Space and Self.* MIT Press, Cambridge.

Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science*, 2:53–62.

Chrisley, R. (1991). A hybrid architecture for cognitive map construction & use. *Artificial Intelligence & the Simulation of Behaviour: Special Issue on Hybrid Models of Cognition*. No. 78.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78(2).

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distibuted Processing*. MIT Press, Cambridge, Mass.

Cleeremans, A. (1992). *Mechanisms of Implicit Learning*. MIT Press, Cambridge.

Crane, T. (1992). The non-conceptual content of experience. In Crane, T., editor, *The Contents of Experience*. Cambridge University Press, Cambridge.

Cussins, A. (1986). *A Representational Theory of Mind*. University of Oxford. D.Phil. thesis.

Cussins, A. (1987). Varieties of psychologism. *Synthese*, 70:123–54.

Cussins, A. (1990). The connectionist construction of concepts. In Boden, M., editor, *The Philosophy of Artificial Intelligence*, pages 368–440. Oxford University Press, Oxford.

Cussins, A. (1992). The limitations of pluralism. In Charles, D. and Lennon, K., editors, *Reduction, Explanation and Realism*, pages 179–223. Clarendon Press, Oxford.

Davidson, D. (1974). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47.

Davies, M. (1990). Thinking persons and cognitive science. *AI and Society*.

Davies, M. (1991). Externalism and perceptual content. *Proceedings of the Aristotelian society*.

Dennett, D. (1987). *The Intentional Stance.* MIT Press, Cambridge, Mass.

Diamond, A. (1988). Differences between adult and infant cognition: Is the crucial variable presence or absence of language? In Weiskrantz, L., editor, *Thought Without Language.* Oxford University Press, Oxford.

Dienes, Z. (1994). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. Lecture given to PDP Discussion Group, School of Cognitive & Computing Sciences, University of Sussex, 18 November 1994.

Dummett, M. (1981). *The Interpretation of Frege's Philosophy.* Duckworth, London.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–212.

Elton, M. (1995). *Presence of mind : a study of consciousness.* University of Sussex, Brighton. D.Phil. thesis.

Evans, G. (1982). *The Varieties of Reference.* Oxford University Press, Oxford.

Evans, G. (1985). Things without the mind. In Phillips, A., editor, *Collected papers.* Oxford University Press, Oxford.

Fahlman, S. (1988). Faster-learning variations on back-propagation: an empirical study. In Touretzky, D., Hinton, G., and Sejnowski, T., editors, *The Proceedings of the 1988 Connectionist Models Summer School*, pages 11–20, San Mateo. Morgan Kaufmann.

Fodor, J. (1981). Methodological solipsism considered as a research strategy in cognitive science. In Haugeland, J., editor, *Mind design*, pages 307–338. MIT Press, Cambridge.

Fodor, J. (1985). Fodor's guide to mental representation – the intelligent auntie's vade-mecum. *Mind*, 94(373):76–100.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* MIT Press, Cambridge, Mass.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In Pinker, S. and Mehler, J., editors, *Connections and Symbols.* MIT Press, Cambridge, Mass.

Glass, A. and Holyoak, K. (1986). *Cognition.* Random House, New York. 2nd edition.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Harris, P. (1989). Object permanence in infancy. In Slater, A. and Bremner, G., editors, *Infant Development*, pages 102–121. Lawrence Erlbaum, Hove.

Harris, P. (1991). The work of the imagination. In Whiten, A., editor, *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Meaning.* Basil Blackwell, Oxford.

Haugeland, J. (1991). Representational genera. In Ramsey, W., Stich, S., and Rumelhart, D., editors, *Philosophy and Connectionist Theory.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Hinton, G., McClelland, J., and Rumelhart, D. (1986). Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 77–109. MIT Press, Cambridge, Mass.

Holland, A. (1994). *Simple Recurrent Networks, Non-conceptual Content and the Development of Objectivity.* University of Sussex, Brighton. MSc thesis.

Hood, B. and Willatts, P. (1986). Reaching in the dark to an object's remembered position: Evidence for object permanence in 5-month-old infants. *British Journal of Developmental Psychology*, 4:57–66.

Langton, C. (1989). Preface. In *Artificial Life: Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, Los Alamos, NM, Sept, 1987.* Addison–Wesley. Volume VI in the series of the Santa Fe Institute Studies in the Sciences of Complexity.

Lloyd Morgan, C. (1894). *Introduction to Comparative Psychology.* Scott, London.

Ludwig, K. (1994). Blueprint for a science of mind: A critical notice of Christopher Peacocke's "A study of concepts". *Mind and Language*, 9(4).

Luria, A. (1968). *The Mind of a Mnemonist.* Basic Books, New York.

Mareschal, D. and Plunkett, K. (1994). Object permanence and visual tracking: A connectionist perspective. In *The Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.*

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In Haugeland, J., editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, pages 143–60. MIT Press, Cambridge, Mass.

McDowell, J. (1987). In defense of modesty. In Taylor, B., editor, *Michael Dummett: Contributions to Philosophy.* Nijhoff, Dordrecht.

McDowell, J. (1994a). The content of perceptual experience. *The Philosophical Quarterly*, 44(175):190–205.

McDowell, J. (1994b). *Mind and World.* Harvard University Press, Cambridge.

McGinn, C. (1982). The structure of content. In Woodfield, A., editor, *Thought and Object: Essays on Intentionality.* Clarendon Press, Oxford.

McGinn, C. (1989). *Mental Content.* Basil Blackwell, Oxford.

Morris, M. (1992). *The Good and the True.* Clarendon Press, Oxford.

Müller-Lyer, F. (1981). Optical illusions. *Perception*, 10(2). Originally published in 1889; translated by R. Day and H. Knuth.

Nagel, T. (1986). *The View from Nowhere.* Oxford University Press, Oxford.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map.* Clarendon Press, Oxford.

Peacocke, C. (1981). Demonstrative thought and psychological explanation. *Synthese*, 49:187–217.

Peacocke, C. (1983). *Sense and Content: Experience, Thought and their Relations.* Clarendon Press, Oxford.

Peacocke, C. (1986). *Thoughts: An Essay on Content.* Basil Blackwell, Oxford.

Peacocke, C. (1989). *Transcendental Arguments in the Theory of Content.* Clarendon Press, Oxford.

Peacocke, C. (1990). Perceptual content. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes from Kaplan.* Oxford University Press, New York.

Peacocke, C. (1992). Scenarios, contents & perception. In Crane, T., editor, *The Contents of Experience.* Cambridge University Press, Cambridge.

Peacocke, C. (1993). *A Study of Concepts.* MIT Press, Cambridge, Mass.

Peacocke, C. (1994a). Content, computation and externalism. *Mind and language*, 9(3):303–335.

Peacocke, C. (1994b). Rationality, norms and the primitively compelling: A reply to Kirk Ludwig. *Mind and Language*, 9(4).

Perry, J. (1979). The problem of the essential indexical. *Nous*, 13:3–21.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence*, 46.

Putnam, H. (1975). The meaning of meaning. In Putnam, H., editor, *Mind, Language and Reality: Philosophical Papers, Volume 2.* Cambridge University Press, Cambridge.

Putnam, H. (1988). *Representation and Reality.* MIT Press, Cambridge, Mass.

Quine, W. (1960). *Word and Object.* MIT press, Cambridge.

Ramsey, W., Stich, S., and Garon, J. (1991). Connectionism, eliminativism, and the future of folk-psychology. In Ramsey, W., Stich, S., and Rumelhart, D., editors, *Philosophy and Connectionist Theory.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Rosenbloom, P., Laird, J., Newell, A., and McCarl, R. (1992). A preliminary analysis of the SOAR achitecture as a basis for general intelligence. In Kirsh, D., editor, *Foundations of Artificial Intelligence*, pages 289–326. MIT Press, Cambridge, Mass.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by back-propagating errors. *Nature*, 323:533–536.

Rutkowska, J. (1993). *The Computational Infant: Looking for Developmental Cognitive Science*. Harvester Wheatsheaf, London.

Searle, J. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64.

Searle, J. (1992). *The Rediscovery of the Mind*. MIT Press, Cambridge, Mass.

Smith, B. (1992). The owl and the electric encyclopedia. In Kirsh, D., editor, *Foundations of Artificial Intelligence*, pages 251–288. MIT Press, Cambridge.

Smith, B. C. (1991). On the threshold of belief. In Kirsh, D., editor, *Foundations of artificial intelligence*. MIT Press, Cambridge.

Smith, B. C. (1996). *On the Origin of Objects*. MIT Press, Cambridge.

Smolensky, P. (1987). On variable binding and the representation of symbolic structures in connectionist systems. Technical Report 355-87, Department of Computer Science, University of Colorado.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Spelke, E. (1985). Perception of unity, persistence, and identity: Thoughts on infants' conceptions of objects. In Mehler, J. and Fox, R., editors, *Neonate Cognition: Beyond the Blooming, Buzzing Confusion*. Lawrence Erlbaum Associates Ltd., London.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. MIT Press, Cambridge.

Strawson, P. (1959). *Individuals*. Methuen, London.

Touretzky, D. and Hinton, G. (1988). A distributed connectionist production system. *Cognitive Science*, 12:423–466.

Travis, C. (1994). On constraints of generality. *Proceedings of the Aristotelian Society*, 94:165–188.

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, 14:355–384.

van Gelder, T. (1991). What is the 'D' in 'PDP'? In Ramsey, W., Stich, S., and Rumelhart, D., editors, *Philosophy and Connectionist Theory*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Woodfield, A. (1993). Do your concepts develop? In Hookway, C. and Peterson, D., editors, *The Proceedings of the 1992 Royal Institute of Philosophy Conference on Philosophy and the Cognitive Sciences*, volume 34, Cambridge. Cambridge University Press. Supplement to *Philosophy*.